GENERALIZED SINGULAR VALUE DECOMPOSITION OF GENOMIC PROFILES PREDICTS GLIOBLASTOMA SURVIVAL

by

Benjamin O. Alpert

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Bioengineering

The University of Utah

August 2012

STATEMENT OF THESIS APPROVAL

The thesis of **Benjamin O. Alpert**

has been approved by the following supervisory committee members:

Orly Alter	, Chair	**7/2/2012** Date Approved
Rob MacLeod	, Member	**6/14/2012** Date Approved
Andrea H. Bild	, Member	**6/22/2012** Date Approved

and by **Patrick A. Tresco** , Chair of

the Department of **Bioengineering**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

Despite recent large-scale profiling efforts, the best prognostic predictor of glioblastoma multiforme (GBM) remains the patient's age at diagnosis. We describe a global pattern of tumor-exclusive co-occurring copy-number alterations (CNAs) that is correlated, possibly coordinated with GBM patients' survival and response to chemotherapy. The pattern is revealed by GSVD comparison of patient-matched but probe-independent GBM and normal aCGH datasets from The Cancer Genome Atlas (TCGA). We find that, first, the GSVD, formulated as a framework for comparatively modeling two composite datasets, removes from the pattern copy-number variations (CNVs) that occur in the normal human genome (e.g., female-specific X chromosome amplification) and experimental variations (e.g., in tissue batch, genomic center, hybridization date and scanner), without a-priori knowledge of these variations. Second, the pattern includes most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs in >3% of the patients. These include the biochemically putative drug target, cell cycle-regulated serine/threonine kinase-encoding *TLK2*, the cyclin E1-encoding *CCNE1*, and the Rb-binding histone demethylase-encoding *KDM5A*. Third, the pattern provides a better prognostic predictor than the chromosome numbers or any one focal CNA that it identifies, suggesting that the GBM survival phenotype is an outcome of its global genotype. The pattern is independent of age, and combined with age, makes a better predictor than age alone. GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set, confirms the global pattern. GSVD classification of the GBM profiles of an independent set of patients validates the prognostic contribution of the pattern.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Orly Alter, for her guidance and contribution to this research study. I also thank my coauthors and friends, Cheng Lee and Preethi Sankaranarayanan.

I thank the members of my thesis committee, Dr. Rob MacLeod and Dr. Andrea Bild, for their valuable insights. Also thanks to Dr. Chris Johnson, Dr. Michael Saunders and Dr. Charles Van Loan for insightful discussions of matrix and tensor computations, and Dr. Howard Colman, Dr. Randy Jensen and Dr. Joshua Schiffman for helpful discussions of glioblastoma cancer genomics. Thanks to Andy Gross for technical assistance.

Financial support for this work was provided by the Utah Science Technology and Research initiative, National Human Genome Research Institute R01 Grant HG-004302, and National Science Foundation CAREER Award DMS-0847173, and is gratefully acknowledged.

CHAPTER 1

BACKGROUND AND INTRODUCTION[1]

1.1 Background

The announcement of the first complete sequencing of the human genome in 2000 was a major milestone that opened a new era in the field of human genomics. It brought promise of real impact on people's lives, and it was expected to revolutionize the diagnosis, prevention and treatment of virtually all human diseases, especially cancer. While there is no denying that genomic research is having a profound impact on scientific progress, the clinical impact so far has been modest.

DNA microarrays are commonly used for analyzing the human genome, estimated to be over 3 billion base pairs long and to contain over 20,000 distinct genes. Microarrays can be used to detect copy-number variations (CNVs), the phenomenon where certain DNA segments are either duplicated or deleted. When a CNV is associated with a pathogenic state, it is considered a copy-number alteration (CNA). Array comparative genomic hybridization (aCGH) is a microarray-based technique for detecting CNAs at a high resolution. The technique includes a test DNA sample and a reference DNA control sample, where the test sample is labeled with a red fluorescent dye, and the control sample is labeled with a green fluorescent dye. The samples are combined in equal amounts and hybridized to probes on the microarray. These probes have short DNA sequences representing specific locations along the genome. Microarray resolution is improving rapidly, and newer platforms have up to 1 million probes covering short (∼60) base pair segments. The microarray is then scanned and the ratio of the test and control DNA are measured in all probes. CNAs in the test sample appear as amplifications (red) or deletions (green) of a certain genomic segment (Figure 1.1).

---

[1]Reprinted with minor revisions and with permission from *Public Library of Science (PLoS) One* 7 (1), article e30098 (January 2012); C. H. Lee,* B. O. Alpert,* P. Sankaranarayanan and O. Alter, *"GSVD Comparison of Patient-Matched Normal and Tumor aCGH Profiles Reveals Global Copy-Number Alterations Predicting Glioblastoma Multiforme Survival"*; http://dx.doi.org/10.1371/journal.pone.0030098.

*These authors contributed equally to this work.

Cancer genomics research has shown that a cancer of a certain type is genetically distinct from patient to patient. For example, clustering analysis of gene expression profiles may distinguish cancer subtypes based on different gene expression signatures. Identifying and understanding patterns of genetic alterations that drive tumor development in each case will allow scientists to better classify tumors. Information about aberrant functional genes will ultimately allow for clinical implementation of gene-targeted therapies that are tailored to patients who are most likely to benefit from these therapies. This information will also improve prognosis prediction which can guide physicians towards the most appropriate treatment. For instance, it has been shown that ovarian cancer patients with *BRCA2* mutations exhibited increased sensitivity to certain chemotherapeutic agents, such as cisplatin [1].

In 2005, The National Cancer Institute and the National Human Genome Research Institute launched The Cancer Genome Atlas (TCGA), a billion-dollar, comprehensive project for cataloguing genetic aberrations associated with cancer, using multiple genome analysis techniques in large patient cohorts. The project aims to collect large numbers of high quality tumor and patient-matched normal samples from over 20 cancers. Some of the techniques used include gene expression profiling, copy-number variation profiling, single nucleotide polymorphisms genotyping, genome-wide DNA methylation profiling, microRNA profiling, and exon sequencing. These data are freely provided to the research community.

The first cancer studied by TCGA is glioblastoma multiforme (GBM), the most common primary brain tumor in adults. GBM arises from the glial cells, which provide support and protection for the neurons. The cancer is usually detected at a late stage after symptoms begin to appear, and it is characterized by poor prognosis [2] with patients having a median survival time of approximately 1 year from the time of diagnosis. GBM tumors exhibit a range of CNAs, many of which play roles in the cancer's pathogenesis [3–5]. Recent large-scale gene expression [6–8] and DNA methylation [9] profiling efforts identified GBM molecular subtypes, distinguished by small numbers of biomarkers. However, despite these efforts, GBM's best prognostic predictor remains the patient's age at diagnosis [10, 11]. Most GBM studies that have found various prognostic biomarkers did so analyzing gene expression and methylation data. In this work, we analyze copy-number variation in tumor and patient-matched normal profiles and identify a pattern that predicts GBM patients' survival.

The advent of new genomic technologies that produce massive datasets has raised a big challenge: How can we store such large outputs, let alone analyze them? The growth in raw

output has outstripped Moore's Law of advances in information technology and storage capacity [12]. This is one of the reasons researchers resort to analyzing smaller subsets of the data, for example choosing genes that are suspected to play roles in the biological processes under investigation. The singular value decomposition (SVD) is a mathematical factorization that can reveal significant patterns in large datasets, and has many useful applications in signal processing and statistics. It is an effective method of reducing the dimensionality of large-scale datasets which include multiple genomic samples, and it has been previously used to model DNA microarray data [13].

TCGA provides cancer genomic data using multiple genome analysis techniques in large patient cohorts. The structure of these data integrated from different studies is of an order higher than that of a matrix, and there is a fundamental need for mathematical frameworks suitable for analyzing such large-scale multidimensional data. The development of tools capable of effectively analyzing these data will help in obtaining biologically meaningful results, which may translate into clinically relevant information, such as aberrant genes and patient prognosis.

One such tool for analyzing multidimensional data is the generalized SVD (GSVD), which can be used to integrate two large-scale matrices of different numbers of rows and the same numbers of columns. Unlike existing algorithms, the GSVD does not require a mapping across the different datasets, allowing conservation of a more full range of the data. It was demonstrated using the GSVD that modeling of DNA microarray data can correctly predict previously unknown cellular mechanisms [14]. GSVD comparative modeling of cancer genomic data, therefore, draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.

1.2 Introduction

To identify CNAs that might predict GBM patients' survival, we comparatively model patient-matched GBM and normal aCGH profiles from TCGA by using the GSVD [15]. Previously, we formulated the GSVD as a framework for comparatively modeling two composite datasets [16] (see also [17]), and illustrated its application in sequence-independent comparison of DNA microarray data from two organisms, where, as we showed, the mathematical variables and operations of the GSVD represent experimental or biological reality. The variables, subspaces of significant patterns that are uncovered in the simultaneous decomposition of the two datasets and are mathematically significant in either both (i.e., common to both) datasets or only one (i.e., exclusive to one) of the datasets, correlate with

cellular programs that are either conserved in both or unique to only one of the organisms, respectively. The operation of reconstruction in the subspaces that are mathematically common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism.

We now find that also in probe-independent comparison of aCGH data from patient-matched tumor and normal samples, the mathematical variables of the GSVD, i.e., shared tumor and normal patterns of copy-number variation across the patients and the corresponding tumor- and normal-specific patterns of copy-number variation across the tumor and normal probes, represent experimental or biological reality. Patterns that are mathematically significant in both datasets represent CNVs in the normal human genome that are conserved in the tumor genome (e.g., female-specific X chromosome amplification). Patterns that are mathematically significant in the normal but not the tumor dataset represent experimental variations that exclusively affect the normal dataset. Similarly, some patterns that are mathematically significant in the tumor but not the normal dataset represent experimental variations that exclusively affect the tumor dataset.

One pattern that is mathematically significant in the tumor but not the normal dataset, represents tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs in >3% of the patients [18]. This pattern is correlated, possibly coordinated with GBM patients' survival and response to therapy. We find that the pattern provides a prognostic predictor that is better than the chromosome numbers or any one focal CNA that it identifies, suggesting that the GBM survival phenotype is an outcome of its global genotype. The pattern is independent of age, and combined with age, makes a better predictor than age alone.

We confirm our results with GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set. We validate the prognostic contribution of the pattern with GSVD classification of the GBM profiles of a set of patients that is independent of both the initial set and the inclusive confirmation set [19].

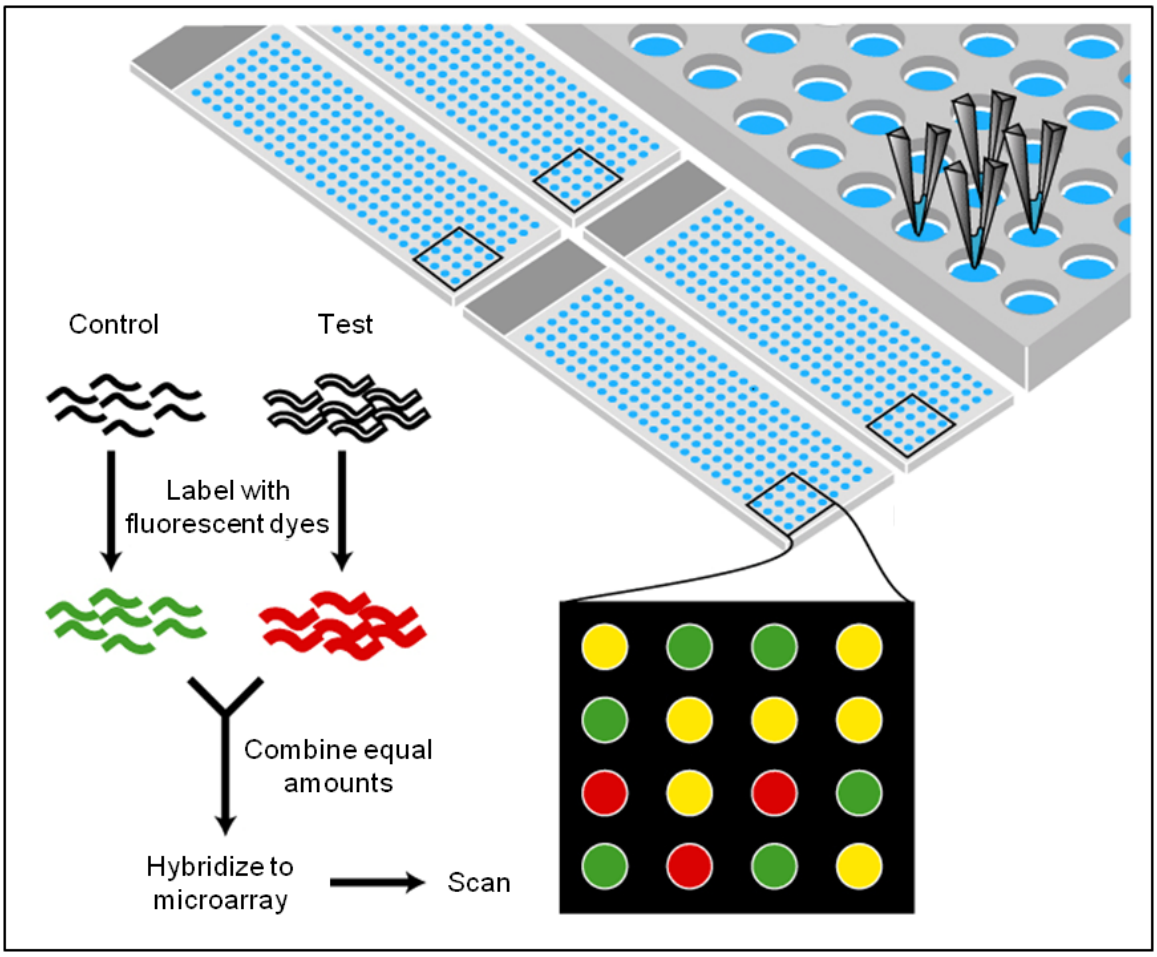


Figure 1.1. Array comparative genomic hybridization (aCGH).

CHAPTER 2

METHODS

To compare TCGA patient-matched GBM and normal (mostly blood) aCGH profiles (Dataset S1 and Mathematica Notebooks S1 and S2), Agilent Human aCGH 244A-measured 365 tumor and 360 normal profiles were selected, corresponding to the same N=251 patients. Each profile lists $\log_2$ of the TCGA level 1 background-subtracted intensity in the sample relative to the Promega DNA reference, with signal to background >2.5 for both the sample and reference in more than 90% of the 223,603 autosomal probes on the microarray. The profiles are organized in one tumor and one normal dataset, of M_1=212,696 and M_2=211,227 autosomal and X chromosome probes, each probe with valid data in at least 99% of either the tumor or normal arrays, respectively. Each profile is centered at its autosomal median copy number. The <0.2% missing data entries in the tumor and normal datasets are estimated by using singular value decomposition (SVD) as described [13, 16, 20–22]. Within each set, the medians of profiles of samples from the same patient are taken.

The structure of the patient-matched but probe-independent tumor and normal datasets D_1 and D_2, of N patients, i.e., N-arrays $\times$ M_1-tumor and M_2-normal probes, is of an order higher than that of a single matrix. The patients, the tumor and normal probes as well as the tissue types, each represent a degree of freedom. Unfolded into a single matrix, some of the degrees of freedom are lost, limiting the possible interpretations of the data (see also [23, 24].

To compare the tumor and normal datasets, therefore, we use the GSVD, formulated to simultaneously separate the paired datasets into paired weighted sums of N outer products of two patterns each: One pattern of copy-number variation across the patients, i.e., a "probelet" v_n^T, which is identical for both the tumor and normal datasets, combined with either the corresponding tumor-specific pattern of copy-number variation across the tumor probes, i.e., the "tumor arraylet" $u_{1,n}$, or the corresponding normal-specific pattern across the normal probes, i.e., the "normal arraylet" $u_{2,n}$ (Figures 2.1 and 2.2),

$$
\begin{aligned}
D_1 &= U_1 \Sigma_1 V^T = \sum_{n=1}^{N} \sigma_{1,n} u_{1,n} \otimes v_n^T, \\
D_2 &= U_2 \Sigma_2 V^T = \sum_{n=1}^{N} \sigma_{2,n} u_{2,n} \otimes v_n^T.
\end{aligned} \tag{2.1}
$$

The probelets are, in general, nonorthonormal, but are normalized, such that $v_n^T v_n = 1$. The tumor and normal arraylets are orthonormal, such that $U_1^T U_1 = U_2^T U_2 = I$.

The significance of the probelet v_n^T in either the tumor or normal dataset, in terms of the overall information that it captures in this dataset, is proportional to either of the weights $\sigma_{1,n}$ or $\sigma_{2,n}$, respectively (Figure 2.3),

$$
\begin{aligned}
p_{1,n} &= \sigma_{1,n}^2 / \sum_{n=1}^{N} \sigma_{1,n}^2, \\
p_{2,n} &= \sigma_{2,n}^2 / \sum_{n=1}^{N} \sigma_{2,n}^2.
\end{aligned} \tag{2.2}
$$

The "generalized normalized Shannon entropy" of each dataset,

$$
\begin{aligned}
0 \leq d_1 &= (\log N)^{-1} \sum_{n=1}^{N} p_{1,n} \log p_{1,n} \leq 1, \\
0 \leq d_2 &= (\log N)^{-1} \sum_{n=1}^{N} p_{2,n} \log p_{2,n} \leq 1,
\end{aligned} \tag{2.3}
$$

measures the complexity of the data from the distribution of the overall information among the different probelets and corresponding arraylets. An entropy of zero corresponds to an ordered and redundant dataset in which all the information is captured by a single probelet and its corresponding arraylet. An entropy of one corresponds to a disordered and random dataset in which all probelets and arraylets are of equal significance. The significance of the probelet v_n^T in the tumor dataset relative to its significance in the normal dataset is defined in terms of an "angular distance" θ_n that is proportional to the ratio of these weights,

$$
-\pi/4 \leq \theta_n = \arctan(\sigma_{1,n}/\sigma_{2,n}) - \pi/4 \leq \pi/4. \tag{2.4}
$$

An angular distance of $\pm\pi/4$ indicates a probelet that is exclusive to either the tumor or normal dataset, respectively, whereas an angular distance of zero indicates a probelet that is common to both the tumor and normal datasets. The probelets are arranged in decreasing order of their angular distances, i.e., their significance in the tumor dataset relative to the normal dataset.

To biologically or experimentally interpret these significant probelets, we correlate or anticorrelate each probelet with relative copy-number gain or loss across a group of patients

according to the TCGA annotations of the group of n patients with largest or smallest relative copy numbers in this probelet among all N patients, respectively. The P-value of a given association is calculated assuming hypergeometric probability distribution of the K annotations among the N patients, and of the subset of $k \subseteq K$ annotations among the subset of n patients, as described [25], $P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{K}{i} \binom{N-K}{n-i}$ (Table 2.1). We visualize the copy-number distribution between the annotations that are associated with largest or smallest relative copy numbers in each probelet by using boxplots, and by calculating the corresponding Mann-Whitney-Wilcoxon P-value (Figures 2.4 and 2.5). To interpret the corresponding tumor and normal arraylets, we map the tumor and normal probes onto the National Center for Biotechnology Information (NCBI) human genome sequence build 36, by using the Agilent Technologies probe annotations posted at the University of California at Santa Cruz (UCSC) human genome browser [26,27]. We segment each arraylet and assign each segment a P-value by using the circular binary segmentation (CBS) algorithm as described (Dataset S2) [28, 29]. We find that the significant probelets and corresponding tumor and normal arraylets, as well as their interpretations, are robust to variations in the preprocessing of the data, e.g., in the data selection cutoffs.

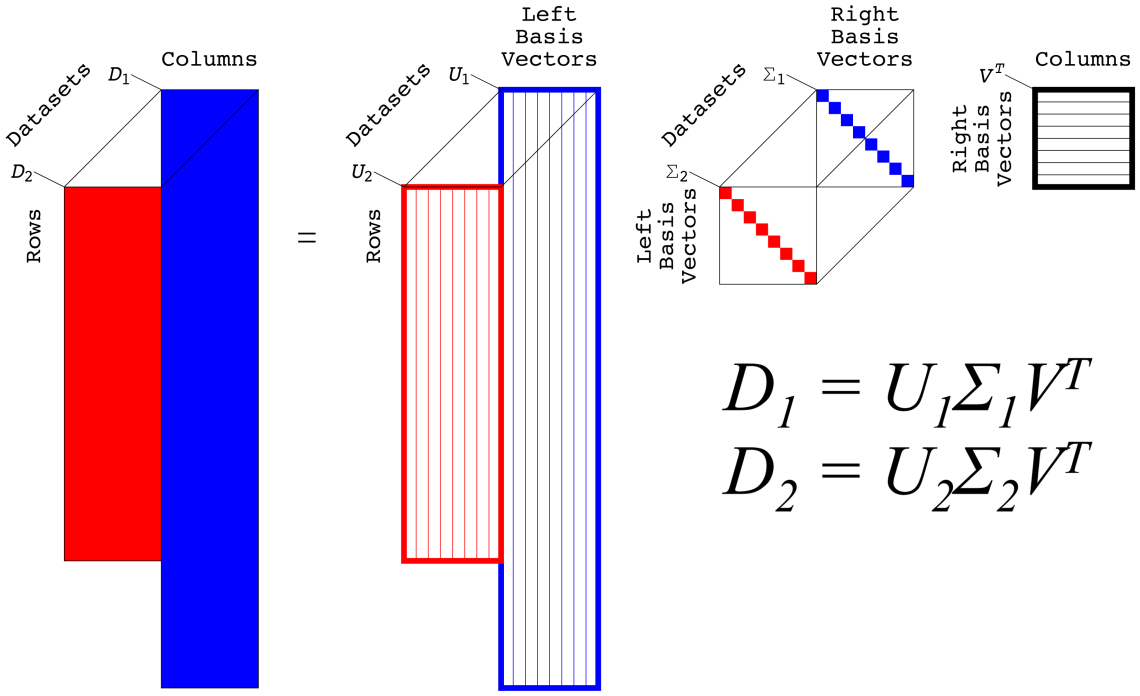
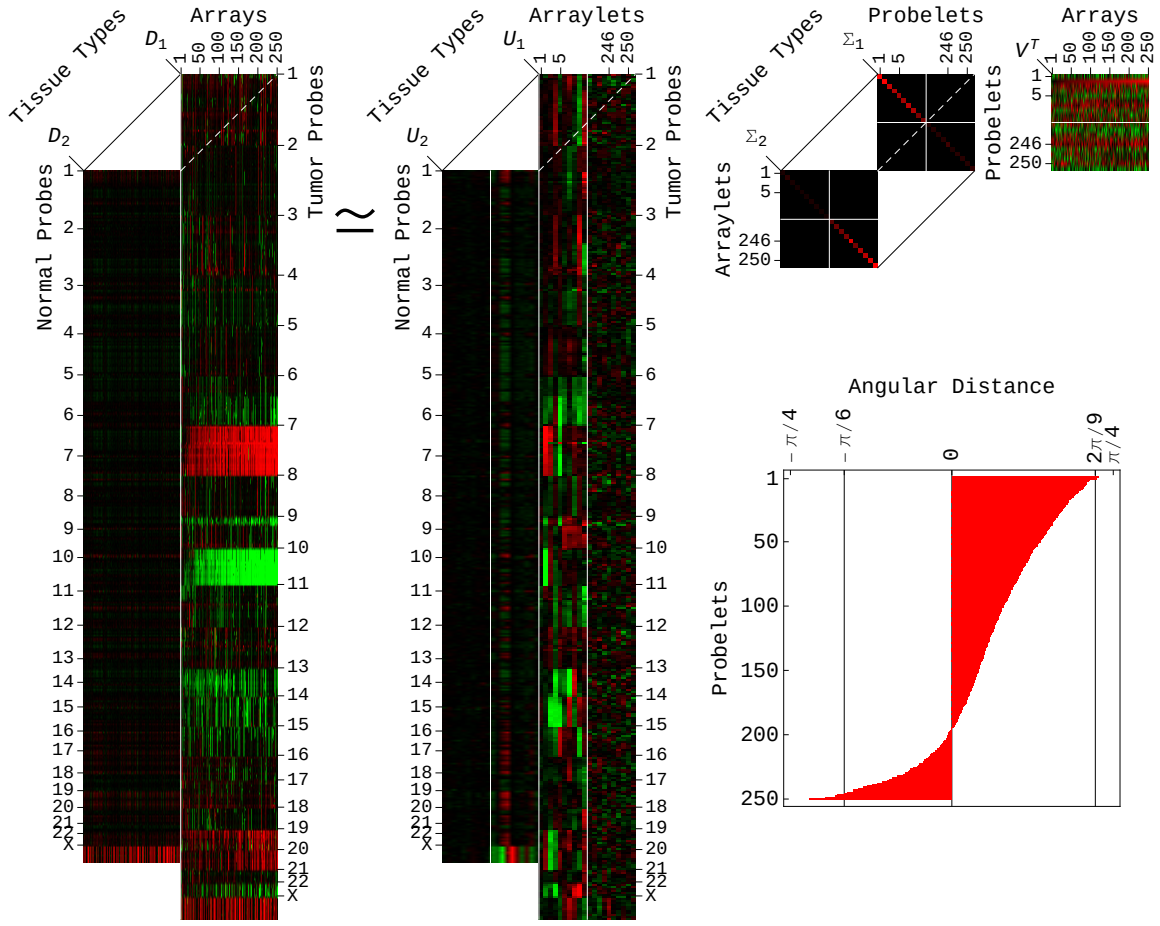


Figure 2.1. Generalized singular value decomposition (GSVD). The data matrices D_1 and D_2 are decomposed into matrices with the same dimensions U_1 and U_2, the square and diagonal matrices Σ_1 and Σ_2, and one shared matrix V^T.

Figure 2.2. Generalized singular value decomposition (GSVD) of the TCGA patient–matched tumor and normal aCGH profiles. The structure of the patient-matched but probe-independent tumor and normal datasets D_1 and D_2, of the initial set of N=251 patients, i.e., N-arrays $\times$ M_1=212,696-tumor probes and M_2=211,227-normal probes, is of an order higher than that of a single matrix. The patients, the tumor and normal probes as well as the tissue types, each represent a degree of freedom. Unfolded into a single matrix, some of the degrees of freedom are lost and much of the information in the datasets might also be lost. The GSVD simultaneously separates the paired datasets into paired weighted sums of N outer products of two patterns each: One pattern of copy-number variation across the patients, i.e., a "probelet" v_n^T, which is identical for both the tumor and normal datasets, combined with either the corresponding tumor-specific pattern of copy-number variation across the tumor probes, i.e., the "tumor arraylet" $u_{1,n}$, or the corresponding normal-specific pattern across the normal probes, i.e., the "normal arraylet" $u_{2,n}$ (Equation 2.1). This is depicted in a raster display, with relative copy-number gain (red), no change (black) and loss (green), explicitly showing only the first through the 10th and the 242nd through the 251st probelets and corresponding tumor and normal arraylets, which capture $\sim$52% and 71% of the information in the tumor and normal dataset, respectively. The significance of the probelet v_n^T in the tumor dataset relative to its significance in the normal dataset is defined in terms of an "angular distance" that is proportional to the ratio of these weights (Equation 2.4). This is depicted in a bar chart display, showing that the first and second probelets are almost exclusive to the tumor dataset with angular distances $>2\pi/9$, the 247th to 251st probelets are approximately exclusive to the normal dataset with angular distances $\lesssim -\pi/6$, and the 246th probelet is relatively common to the normal and tumor datasets with an angular distance $> -\pi/6$. We find and confirm that the second most tumor-exclusive probelet, which is also the most significant probelet in the tumor dataset, significantly correlates with GBM prognosis. The corresponding tumor arraylet describes a global pattern of tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs, including the biochemically putative drug target-encoding *TLK2* [30–33]. We find and validate that a negligible weight of the global pattern in a patient's GBM aCGH profile is indicative of a significantly longer GBM survival time. It was shown that the GSVD provides a mathematical framework for comparative modeling of DNA microarray data from two organisms [16, 22]. Recent experimental results [14] verify a computationally predicted genome-wide mode of regulation [34, 35], and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms. This GSVD comparative modeling of aCGH data from patient-matched tumor and normal samples, therefore, draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.


Tissue Types
Arrays
D_1
D_2
Normal Probes
Tumor Probes
≃
Arraylets
U_1
U_2
Probelets
Σ_1
Σ_2
V^T
Angular Distance
$-\pi/4$
$-\pi/6$
0
$2\pi/9$
$\pi/4$

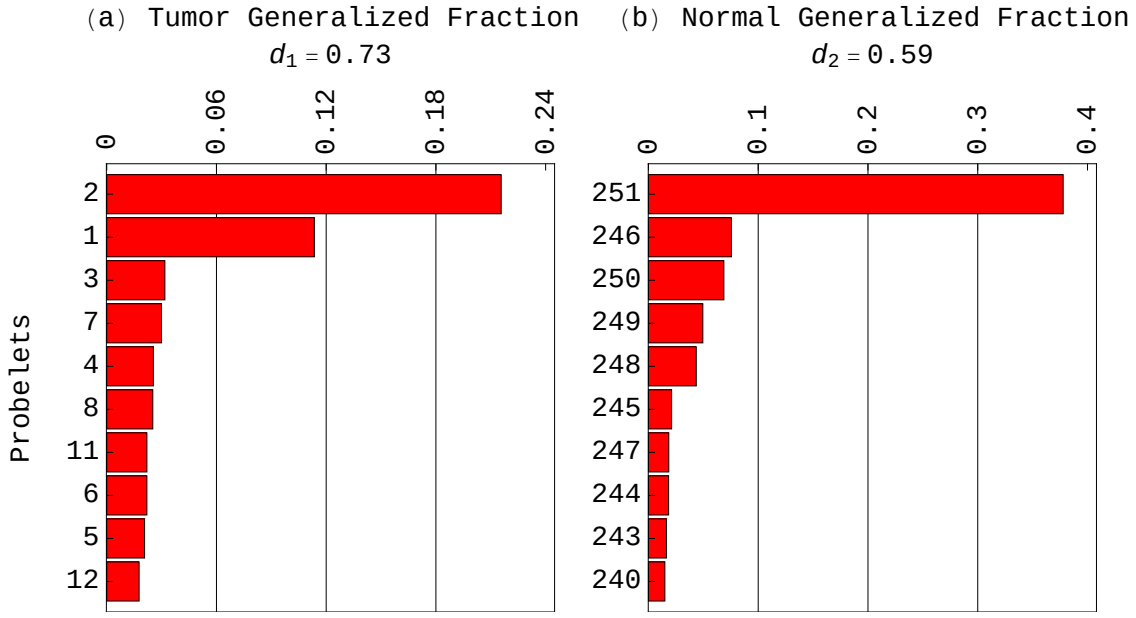


Figure 2.3. Most significant probelets in the tumor and normal datasets. (*a*) Bar chart of the ten most significant probelets in the tumor dataset in terms of the generalized fraction that each probelet captures in this dataset (Equation 2.2), showing that the two most tumor-exclusive probelets, i.e., the first probelet and the second probelet, with angular distances $>2\pi/9$, are also the two most significant probelets in the tumor dataset, with $\sim$11% and 22% of the information in this dataset, respectively. The "generalized normalized Shannon entropy" (Equation 2.3) of the tumor dataset is d_1=0.73. (*b*) Bar chart of the generalized fractions of the ten most significant probelets in the normal dataset, showing that the five most normal-exclusive probelets, the 247th to 251st probelets, with angular distances $\lesssim -\pi/6$, are among the seven most significant probelets in the normal dataset, capturing together $\sim$56% of the information in this dataset. The 246th probelet (Figure 2.2 *d*–*f*), which is relatively common to the normal and tumor datasets with an angular distance $>-\pi/6$, is the second most significant probelet in the normal dataset with $\sim$8% of the information. The generalized entropy of the normal dataset, d_2=0.59, is smaller than that of the tumor dataset. This means that the normal dataset is more redundant and less complex than the tumor dataset.

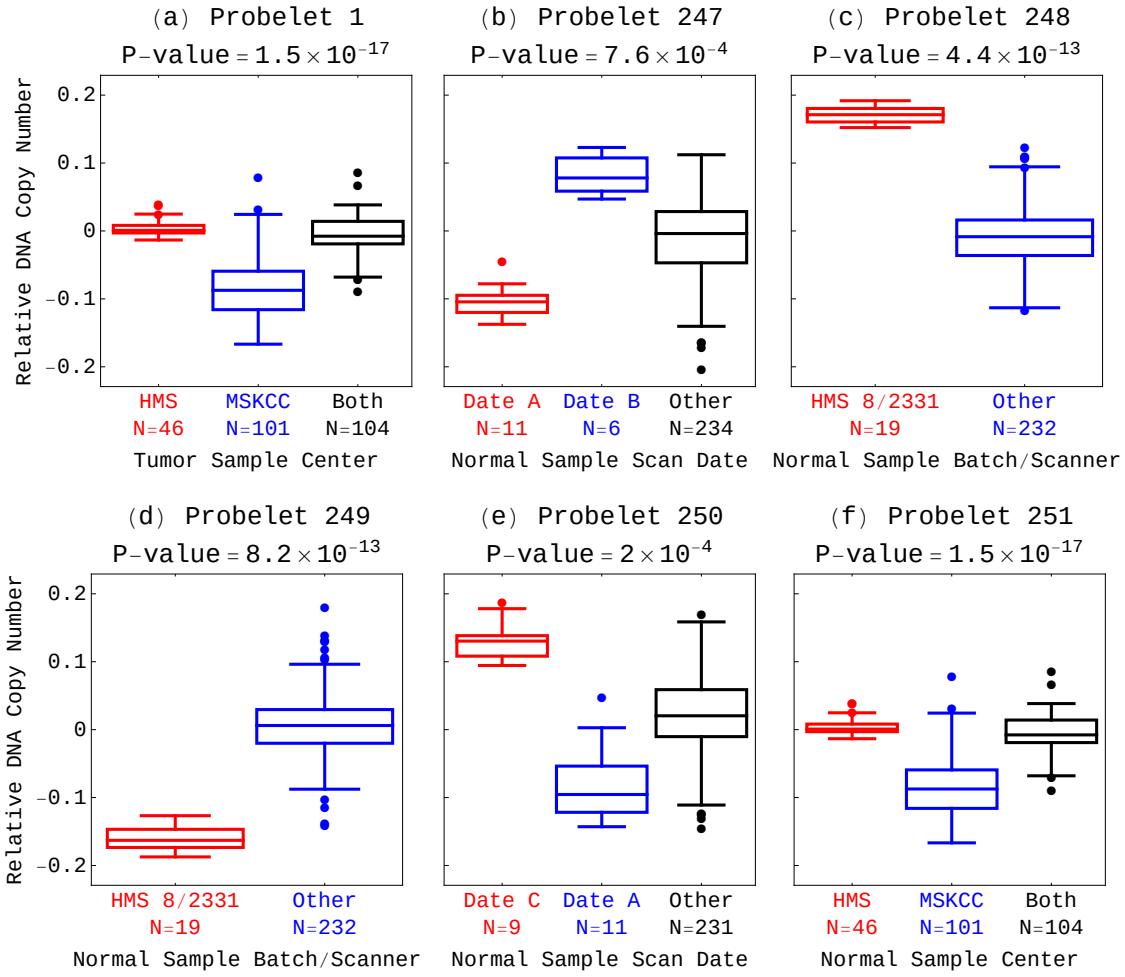


Figure 2.4. Differences in copy numbers among the TCGA annotations associated with the significant probelets. Boxplot visualization of the distribution of copy numbers of the (*a*) first, most tumor-exclusive probelet among the associated genomic centers where the GBM samples were hybridized at (Table 2.1); (*b*) 247th, normal-exclusive probelet among the dates of hybridization of the normal samples; (*c*) 248th, normal-exclusive probelet between the associated tissue batches/hybridization scanners of the normal samples; (*d*) 249th, normal-exclusive probelet between the associated tissue batches/hybridization scanners of the normal samples; (*e*) 250th, normal-exclusive probelet among the dates of hybridization of the normal samples; (*f*) 251st, most normal-exclusive probelet among the associated genomic centers where the normal samples were hybridized at. The Mann-Whitney-Wilcoxon *P*-values correspond to the two annotations that are associated with largest or smallest relative copy numbers in each probelet.

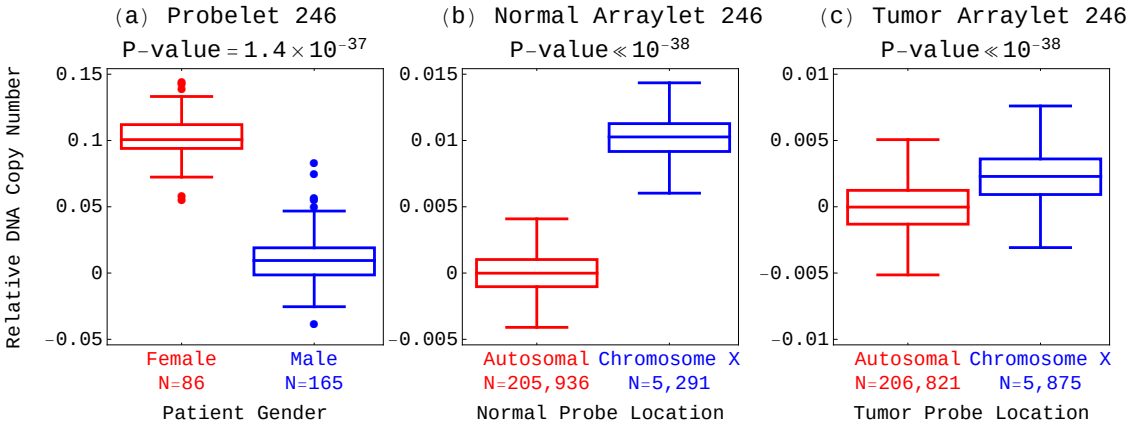


Figure 2.5. Copy-number distributions of the 246th probelet and the corresponding 246th normal arraylet and 246th tumor arraylet. Boxplot visualization and Mann-Whitney-Wilcoxon P-values of the distribution of copy numbers of the (a) 246th probelet, which is approximately common to both the normal and tumor datasets, and is the second most significant in the normal dataset (Figure 2.3b), between the gender annotations (Table 2.1); (b) 246th normal arraylet between the autosomal and X chromosome normal probes; (c) 246th tumor arraylet between the autosomal and X chromosome tumor probes.

Table 2.1. Enrichment of the significant probelets in TCGA annotations. Probabilistic significance of the enrichment of the n patients, with largest or smallest relative copy numbers in each significant probelet, in the respective TCGA annotations. The P-value of each enrichment is calculated assuming hypergeometric probability distribution of the K annotations among the N=251 patients of the initial set, and of the subset of $k \subseteq K$ annotations among the subset of n patients, as described [25], $P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{K}{i} \binom{N-K}{n-i}$.

		Relative DNA Copy Number Gain					Relative DNA Copy Number Loss				
Probelet	Phenotype	Annotation	n	K	k	P-value	Annotation	n	K	k	P-value
1	Tumor Sample Center	HMS	183	34	34	8.5×10^{-6}	MSKCC	68	103	55	3.9×10^{-15}
246	Patient Gender	Female	86	86	84	8.0×10^{-62}	Male	165	165	163	8.0×10^{-62}
247	Normal Sample Scan Date	10.8.2009	51	6	6	5.5×10^{-5}	7.22.2009	38	11	10	2.0×10^{-8}
248	Normal Sample Batch/Scanner	HMS 8/2331	19	19	19	6.2×10^{-29}	–	–	–	–	–
249	Normal Sample Batch/Scanner	–	–	–	–	–	HMS 8/2331	22	19	19	9.6×10^{-26}
250	Normal Sample Scan Date	4.18.2007	26	9	9	3.3×10^{-10}	7.22.2009	25	11	9	1.1×10^{-8}
251	Normal Sample Center	HMS	139	46	46	2.8×10^{-14}	MSKCC	112	101	89	2.1×10^{-32}

CHAPTER 3

RESULTS

We find that, first, the GSVD identifies significant experimental variations that exclusively affect either the tumor or the normal dataset, as well as CNVs that occur in the normal human genome and are common to both datasets, without a-priori knowledge of these variations. The mathematically most tumor-exclusive probelet, i.e., the first probelet (Figure 3.1), correlates with tumor-exclusive experimental variation in the genomic center where the GBM samples were hybridized at, Harvard Medical School (HMS) or Memorial Sloan-Kettering Cancer Center (MSKCC), with the P-values $< 10^{-5}$ (Table 2.1 and Figure 2.4). Similarly, the five most normal-exclusive probelets, i.e., the 247th to 251st probelets (Figures 3.2–3.6), correlate with experimental variations among the normal samples in genomic center, DNA microarray hybridization or scan date as well as the tissue batch and hybridization scanner, with P-values $< 10^{-3}$. Consistently, the corresponding arraylets, i.e., the first tumor arraylet and the 247th to 251st normal arraylets, describe copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths.

We find that the two most tumor-exclusive mathematical patterns of copy-number variation across the patients, i.e., the first probelet (Figure 3.1) and the second probelet (Figure 3.7 *a*–*c*), with angular distances $> 2\pi/9$, are also the two most significant probelets in the tumor dataset, with ∼11% and 22% of the information in this dataset, respectively. Similarly, the five most normal-exclusive probelets, the 247th to 251st probelets (Figures 3.2–3.6), with angular distances $\lesssim -\pi/6$, are among the seven most significant probelets in the normal dataset, capturing together ∼56% of the information in this dataset. The 246th probelet (Figure 3.7*d*–*f*), which is the second most significant probelet in the normal dataset with ∼8% of the information, is relatively common to the normal and tumor datasets with an angular distance $> -\pi/6$.

The 246th probelet (Figure 3.7*e*), which is mathematically approximately common to both the normal and tumor datasets, describes copy-number amplification in the female

relative to the male patients that is biologically common to both the normal and tumor datasets. Consistently, both the 246th normal arraylet (Figure 3.7*d*) and 246th tumor arraylet describe an X chromosome-exclusive amplification. The P-values are $< 10^{-38}$ (Table 2.1 and Figure 2.5). To assign the patients gender, we calculate for each patient the standard deviation of the mean X chromosome number from the autosomal genomic mean in the patient's normal profile (Figure 3.7*f*). Patients with X chromosome amplification greater than twice the standard deviation are assigned the female gender. For three of the patients, this copy-number gender assignment conflicts with the TCGA gender annotation. Upon contacting TCGA, it was confirmed that the TCGA gender assignment in these three patients was incorrect. For three additional patients, the TCGA gender annotation is missing. In all six cases, the classification of the patients by the 246th probelet agrees with the copy-number assignment.

On June 2011, it was noted by the National Human Genome Research Institute that while genome-wide association studies have identified over 1,200 statistically significant associations, only 7 such associations have been reported on the X chromosome. This is mostly due to the exclusion of the X chromosome data from most computational analyses of large-scale molecular biological data. It is, therefore, of particular significance that our analysis includes all autosomal and X chromosome probes with valid data in at least 99% of either the tumor or normal arrays, respectively.

Second, we find that the GSVD identifies a global pattern of tumor-exclusive co-occurring CNAs that includes most known GBM-associated changes in chromosome numbers and focal CNAs. This global pattern is described by the second tumor arraylet (Figure 3.7*a* and Dataset S3). The second most tumor-exclusive probelet (Figure 3.7*b*), which describes the corresponding copy-number variation across the patients, is the most significant probelet in the tumor dataset. Dominant in the global pattern, and frequently observed in GBM samples [3], is a co-occurrence of a gain of chromosome 7 and losses of chromosome 10 and the short arm of chromosome 9 (9p). To assign a chromosome gain or loss, we calculate for each tumor profile the standard deviation of the mean chromosome number from the autosomal genomic mean, excluding the outlying chromosomes 7, 9p and 10. The gain of chromosome 7 and the losses of chromosomes 10 and 9p are greater than twice the standard deviation in the global pattern as well as the tumor profiles of ∼20%, 41% and 12% of the patients, respectively.

Focal CNAs that are known to play roles in the origination and development of GBM and are described by the global pattern include amplifications of segments containing the

genes *MDM4* (1q32.1), *AKT3* (1q44), *EGFR* (7p11.2), *MET* (7q31.2), *CDK4* (12q14.1) and *MDM2* (12q15), and deletions of segments containing the genes *CDKN2A/B* (9p21.3) and *PTEN* (10q23.31), that occur in >3% of the patients. To assign a CNA in a segment, we calculate for each tumor profile the mean segment copy number. Profiles with segment amplification or deletion greater than twice the standard deviation from the autosomal genomic mean, excluding the outlying chromosomes 7, 9p and 10, or greater than one standard deviation from the chromosomal mean, when this deviation is consistent with the deviation from the genomic mean, are assigned a segment gain or loss, respectively. The frequencies of amplification or deletion we observe for these segments are similar to the reported frequencies of the corresponding focal CNAs [5].

Novel CNAs, previously unrecognized in GBM, are also revealed by the global pattern [18]. These include an amplification of a segment that contains *TLK2* (17q23.2) in ∼22% of the patients, with the corresponding CBS P-value$< 10^{-140}$. Copy-number amplification of *TLK2* has been correlated with overexpression in several other cancers [30, 31]. The human gene *TLK2*, with homologs in the plant *Arabidopsis thaliana* but not in the yeast *Saccharomyces cerevisiae*, encodes for a multicellular organisms-specific serine/threonine protein kinase, a biochemically putative drug target [33], which activity is directly dependent on ongoing DNA replication [32]. On the same segment with *TLK2*, we also find the gene *METTL2A*. Another amplified segment (CBS P-value$< 10^{-13}$) contains the homologous gene *METTL2B* (7q32.1). Overexpression of *METTL2A/B* was linked with prostate cancer metastasis [36], cAMP response element-binding (CREB) regulation in myeloid leukemia [37], and breast cancer patients' response to chemotherapy [38].

An amplification of a segment (CBS P-value$< 10^{-145}$) encompassing the cyclin E1-encoding *CCNE1* (19q12) is revealed in ∼4% of the patients. Cyclin E1 regulates entry into the DNA synthesis phase of the cell division cycle. Copy number increases of *CCNE1* have been linked with multiple cancers [39–41], but not GBM. Amplicon-dependent expression of *CCNE1*, together with the genes *POP4*, *PLEKHF1*, *C19orf12* and *C19orf2* that flank *CCNE1* on this segment, was linked with primary treatment failure in ovarian cancer, possibly due to rapid repopulation of the tumor after chemotherapy [42].

Another rare amplification in ∼4% of the patients, of a segment (CBS P-value$< 10^{-28}$) that overlaps with the 5' end of *KDM5A* (12p13.33), is also revealed. The protein encoded by *KDM5A*, a retinoblastoma tumor suppressor (Rb)-binding lysine-specific histone demethylase [43], has been recently implicated in cancer drug tolerance [44]. The same amplified segment includes the solute carrier (SLC) sodium-neurotransmitter symporters

SLC6A12/13, biochemically putative carriers of drugs that might overcome the blood-brain barrier [45]. On the same segment we also find *IQSEC3*, a mature neuron-specific guanine nucleotide exchange factor for the ADP-ribosylation factor *ARF1*, a key regulator of intracellular membrane traffic [46].

Note that although the tumor samples exhibit female-specific X chromosome amplification (Figure 3.7*c*), the second tumor arraylet exhibits an unsegmented X chromosome copy-number distribution, that is approximately centered at zero with a relatively small width. This illustrates the mathematical separation of the global pattern of tumor-exclusive co-occurring CNAs, that is described by the second tumor arraylet, from all other biological and experimental variations that compose either the tumor or the normal dataset, such as the gender variation that is common to both datasets, and is described by the 246th probelet and the corresponding 246th tumor and 246th normal arraylets.

Third, we find that the GSVD classifies the patients into two groups of significantly different prognoses. The classification is according to the copy numbers listed in the second probelet, which correspond to the weights of the second tumor arraylet in the GBM aCGH profiles of the patients. A group of 227 patients, 224 of which with TCGA annotations, displays high (>0.02) relative copy numbers in the second probelet, and a Kaplan-Meier (KM) [47] median survival time of ∼13 months (Figure 3.8*a*). A group of 23 patients, i.e., ∼10% of the patients, displays low, approximately zero, relative copy numbers in the second probelet, and a KM median survival time of ∼29 months, which is more than twice as long as that of the previous group. The corresponding log-rank test P-value is $< 10^{-3}$. The univariate Cox [48] proportional hazard ratio is 2.3, with a P-value $< 10^{-2}$ (Table 3.1), meaning that high relative copy numbers in the second probelet confer more than twice the hazard of low numbers. Note that the cutoff of ± 0.02 was selected to enable classification of as many of the patients as possible. Only one of the 251 patients has a negative copy number in the second probelet <-0.02, and remains unclassified. This patient is also missing the TCGA annotations. Survival analysis of only the chemotherapy patients classified by GSVD gives similar results (Table 3.2 and Figure 3.9*a*). The P-values are calculated without adjusting for multiple comparisons [49]. We observe, therefore, that a negligible weight of the global pattern in a patient's GBM aCGH profile is indicative of a significantly longer survival time, as well as an improved response to treatment among chemotherapy patients.

A mutation in the gene *IDH1* was recently linked with improved GBM prognosis [2, 7] and associated with a CpG island methylator phenotype [9]. We find, however, only seven

patients (six chmeotherapy patients), i.e., <3%, with *IDH1* mutation. This is less than a third of the 23 patients in the long-term survival group defined by the global pattern. The corresponding survival analyses are, therefore, statistically insignificant (Figures 3.10 and 3.11).

Chromosome 10 loss, chromosome 7 gain and even loss of 9p, which are dominant in the global pattern, have been suggested as indicators of poorer GBM prognoses for over two decades [3, 4]. However, the KM survival curves for the groups of patients with either one of these chromosome number changes almost overlap the curves for the patients with no changes (Figure 3.12). The log-rank test P-values for all three classifications are $\gtrsim 10^{-1}$, with the median survival time differences $\lesssim$3 months. Similarly, in the KM survival analyses of the groups of patients with either a CNA or no CNA in either one of the 130 segments identified by the global pattern (Figure 3.13), log-rank test P-values $< 5 \times 10^{-2}$ are calculated for only 12 of the classifications. Of these, only six correspond to a KM median survival time difference that is $\gtrsim$5 months, approximately a third of the $\sim$16 months difference observed for the GSVD classification.

One of these segments contains the genes *TLK2* and *METTL2A* and another segment contains the homologous gene *METTL2B*, previously unrecognized in GBM. The KM median survival times we calculate for the 56 patients with *TLK2/METTL2A* amplification and, separately, for the 19 patients with *METTL2B* amplifications are $\sim$5 and 8 months longer than that for the remaining patients in each case. Similarly, the KM median survival times we calculate for the 43 chemotherapy patients with *TLK2/METTL2A* amplification and, separately, for the 15 chemotherapy patients with *METTL2B* amplification, are both $\sim$7 months longer than that for the remaining chemotherapy patients in each case (Figure 3.14). This suggests that drug-targeting the kinase that *TLK2* encodes and/or the methyltransferase-like proteins that *METTL2A/B* encode may affect not only the pathogenesis but also the prognosis of GBM as well as the patient's response to chemotherapy.

Taken together, we find that the global pattern provides a better prognostic predictor than the chromosome numbers or any one focal CNA that it identifies. This suggests that the GBM survival phenotype is an outcome of its global genotype.

Despite the recent genome-scale molecular profiling efforts, age at diagnosis remains the best prognostic predictor for GBM in clinical use. The KM median survival time difference between the patients >50 or <50 years old at diagnosis is $\sim$11 months, approximately two thirds of the $\sim$16 months difference observed for the global pattern, with the log-rank test P-value $< 10^{-4}$ (Figure 3.8*b*). The univariate Cox proportional hazard ratio we calculate

for age is 2, i.e., similar to that for the global pattern. Taken together, the prognostic contribution of the global pattern is comparable to that of age. Similarly we find that the prognostic contribution of the global pattern is comparable to that of chemotherapy (Figure 3.15a).

To examine whether the weight of the global pattern in a patient's GBM aCGH profile is correlated with the patient's age at diagnosis, we classify the patients into four groups, with prognosis of longer-term survival according to both, only one or neither of the classifications (Figure 3.8c). The KM curves for these four groups are significantly different, with the log-rank test P-value $< 10^{-4}$. Within each age group, the subgroup of patients with low relative copy numbers in the second probelet consistently exhibits longer survival than the remaining patients. The median survival time of the 16 patients <50 years old at diagnosis with low copy numbers in the second probelet is ∼34 months, almost three times longer than the ∼12 months median survival time of the patients >50 years old at diagnosis with high numbers in the second probelet. The multivariate Cox proportional hazard ratios for the global pattern and age are 1.8 and 1.7, respectively, with both corresponding P-values $< 3 \times 10^{-2}$. This survival model relates the time that passed before a patient dies to the global pattern and age covariates, accounting for censoring (time of last patient follow-up). These ratios are similar, meaning that both a high weight of the global pattern in a patient's GBM aCGH profile and an age >50 years old at diagnosis confer similar relative hazard. These ratios also do not differ significantly from the univariate ratios of 2.3 and 2 for the global pattern and age, respectively. Taken together, the prognostic contribution of the global pattern is not only comparable to that of age, but is also independent of age. Combined with age, the global pattern makes a better predictor than age alone, with a combined hazard ratio of ∼3 (multiplying hazard ratios 1.8 and 1.7). We note that the sample size for the group of patients >50 years old with a low weight of the global pattern is small and consists of only 7 patients, therefore the added prognostic contribution of the global pattern is more significant for the group of patients who are <50 years old. Similarly, we find that the global pattern is independent of chemotherapy (Figure 3.15b).

To confirm the global pattern, we use GSVD to compare matched profiles of a larger, more recent, set of 344 TCGA patients, that is inclusive of the initial set of 251 patients [19]. Agilent Human aCGH 244A-measured 458 tumor and 459 normal profiles were selected, corresponding to the inclusive confirmation set of N=344 patients (Dataset S4). The profiles, centered at their autosomal median copy numbers, are organized in one tumor and one normal dataset, of M_1=200,139 and M_2=198,342 probes, respectively. Within each set,

the medians of profiles of samples from the same patient are taken after estimating missing data by using SVD. We find that the significant probelets and corresponding tumor and normal arraylets, as well as their interpretations, are robust to the increase from 251 patients in the initial set to 344 patients in the inclusive confirmation set, and the accompanying decreases in tumor and normal probes, respectively.

The second tumor arraylet computed by GSVD for the 344 patients of the inclusive confirmation set correlates with that of the initial set, with the correlation ∼0.99. To classify the patients according to the copy numbers listed in the corresponding second probelet of the inclusive confirmation set, the classification cutoff ±0.02 of the initial set of 251 patients is scaled by the norm of the copy numbers listed for these patients, resulting in the cutoff ±0.017. Only four of the 251 patients in the initial set, i.e., ∼1.5%, with copy numbers that are near the classification cutoffs of both sets, change classification. Of the 344 patients, we find that 315 patients, 309 with TCGA annotations, display high (>0.017) and 27, i.e., ∼8%, display low, approximately zero, relative copy numbers in the second probelet. Only two patients, one missing TCGA annotations, remain unclassified with large negative (<-0.017) copy numbers in the second probelet. Survival analyses of the inclusive confirmation set of 344 patients give qualitatively the same results as these of the initial set of 251 patients. These analyses confirm that a negligible weight of the global pattern, which is described by the second tumor arraylet, i.e., a low copy number in the second probelet, is indicative of a significantly longer survival time (Figure 3.8*d*). Survival analysis of only the chemotherapy patients in the inclusive confirmation set classified by GSVD gives similar results (Figure 3.9*b*). These analyses confirm that the prognostic contribution of the global pattern is comparable to that of age (Figure 3.8*e*) and is independent of age (Figure 3.8*f*). Similarly, we confirm that the global pattern is independent of chemotherapy (Figures 3.15 *c* and *d*).

To validate the prognostic contribution of the global pattern, we classify GBM profiles of an independent set of 184 TCGA patients, that is mutually exclusive of the initial set of 251 patients. Agilent Human aCGH 244A-measured 280 tumor profiles were selected, corresponding to the independent validation set of 184 patients with available TCGA status annotations (Dataset S5). Each profile lists relative copy numbers in more than 97.5% of the 206,820 autosomal probes among the M_1=212,696 probes that define the second tumor arraylet computed by GSVD for the 251 patients of the initial set. Medians of profiles of samples from the same patient are taken. To classify the 184 patients according to the correlations of their GBM profiles with the second tumor arraylet of the initial set, the classification cutoff of the initial set of 251 patients is scaled by the norm of the correlations

calculated for these patients, resulting in the cutoff ±0.15. For the profiles of 162 patients we calculate high (>0.15) and for 21, i.e., ∼11%, low, approximately zero, correlation with the second tumor arraylet. One patient remains unclassified with a large negative (<-0.15) correlation.

We find that survival analyses of the independent validation set of 184 patients give qualitatively the same results as these of the initial set of 251 patients and the inclusive confirmation set of 344 patients (Figures 3.8 *g*–*i* and Figures 3.9*c*, and 3.15 *e* and *f*). These analyses validate the prognostic contribution of the global pattern, which is computed by GSVD of patient-matched tumor and normal aCGH profiles, also for patients with measured GBM aCGH profiles in the absence of matched normal profiles.

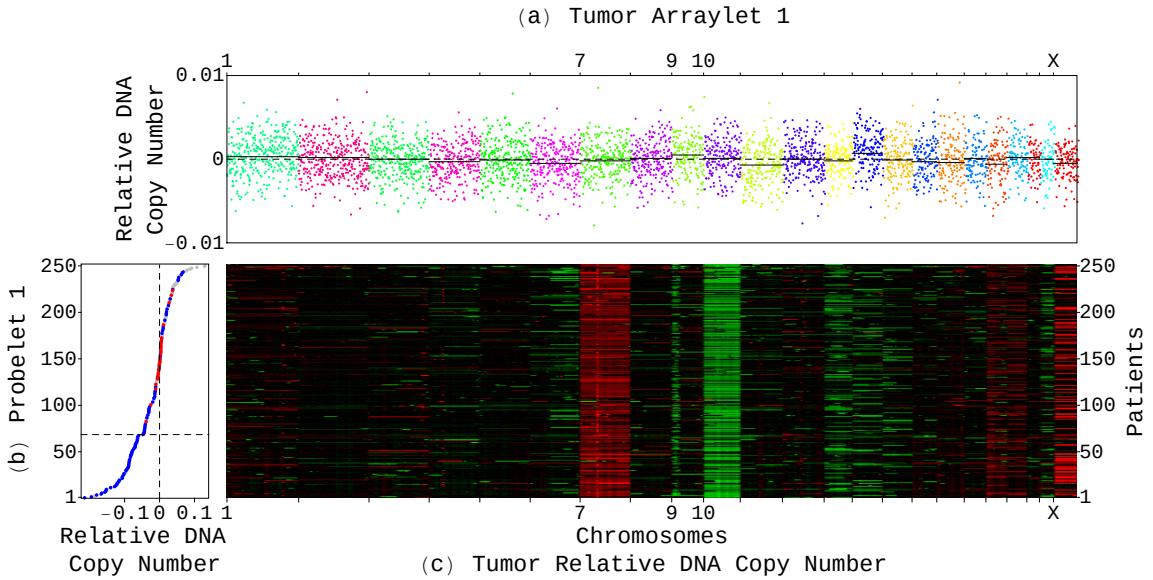


Figure 3.1. The first most tumor-exclusive probelet and corresponding tumor arraylet uncovered by GSVD of the patient-matched GBM and normal aCGH profiles. (*a*) Plot of the first tumor arraylet describes unsegmented chromosomes (black lines), each with copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. (*b*) Plot of the first most tumor-exclusive probelet, which is also the second most significant probelet in the tumor dataset (Figure 2.3*a*), describes the corresponding variation across the patients. The patients are ordered according to each patient's relative copy number in this probelet. These copy numbers significantly correlate with the genomic center where the GBM samples were hybridized at, HMS (red), MSKCC (blue) or multiple locations (gray), with the P-values $<10^{-5}$ (Table 2.1 and Figure 2.4*a*). (*c*) Raster display of the tumor dataset, with relative gain (red), no change (black) and loss (green) of DNA copy numbers, shows the correspondence between the GBM profiles and the first probelet and tumor arraylet.

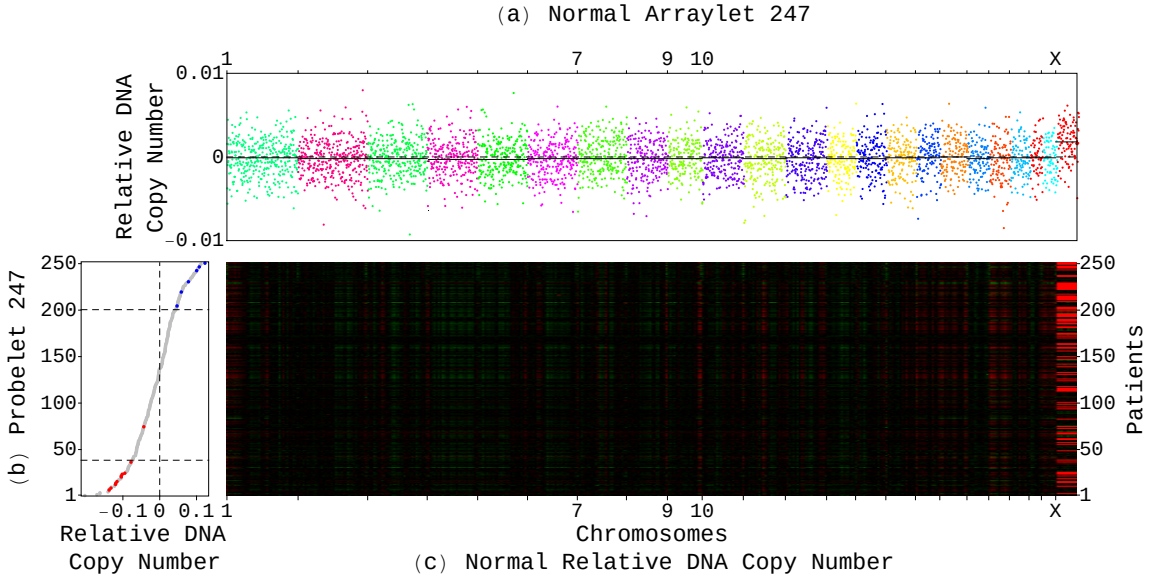


Figure 3.2. The 247th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. (a) Plot of the 247th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. The normal probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. (b) Plot of the 247th probelet describes the corresponding variation across the patients. Copy numbers in this probelet correlate with the date of hybridization of the normal samples, 7.22.2009 (red), 10.8.2009 (blue) or other (gray), with the P-values $<10^{-3}$ (Table 2.1 and Figure 2.4b). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 247th probelet and normal arraylet.

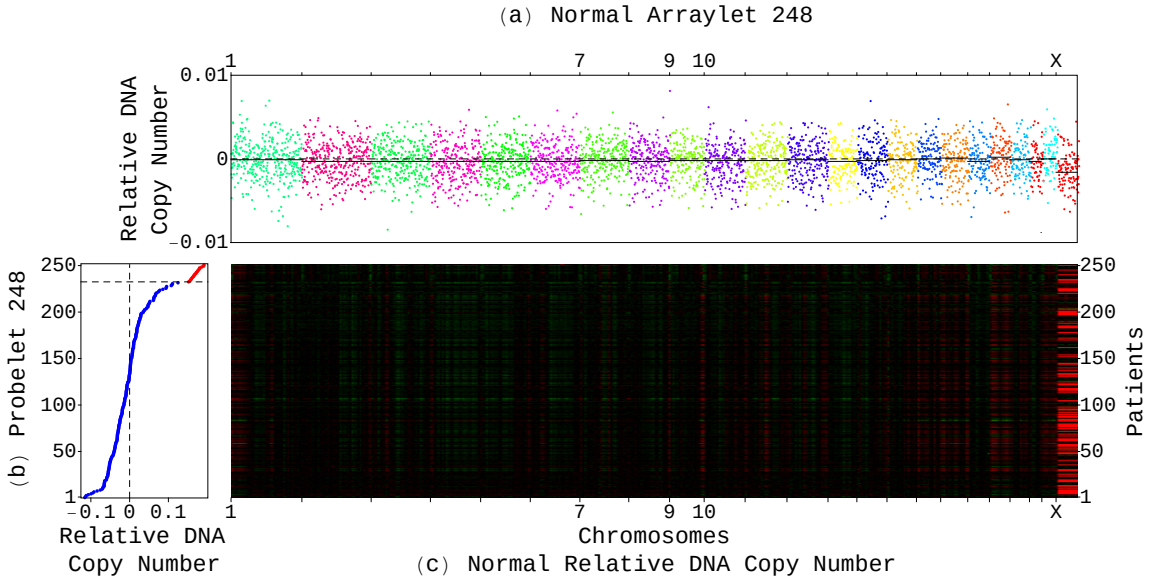


Figure 3.3. The 248th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. (a) Plot of the 248th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the 248th probelet describes the corresponding variation across the patients. Copy numbers in this probelet significantly correlate with the tissue batch/hybridization scanner of the normal samples, HMS 8/2331 (red) and other (gray), with the P-values $<10^{-12}$ (Table 2.1 and Figure 2.4c). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 248th probelet and normal arraylet.

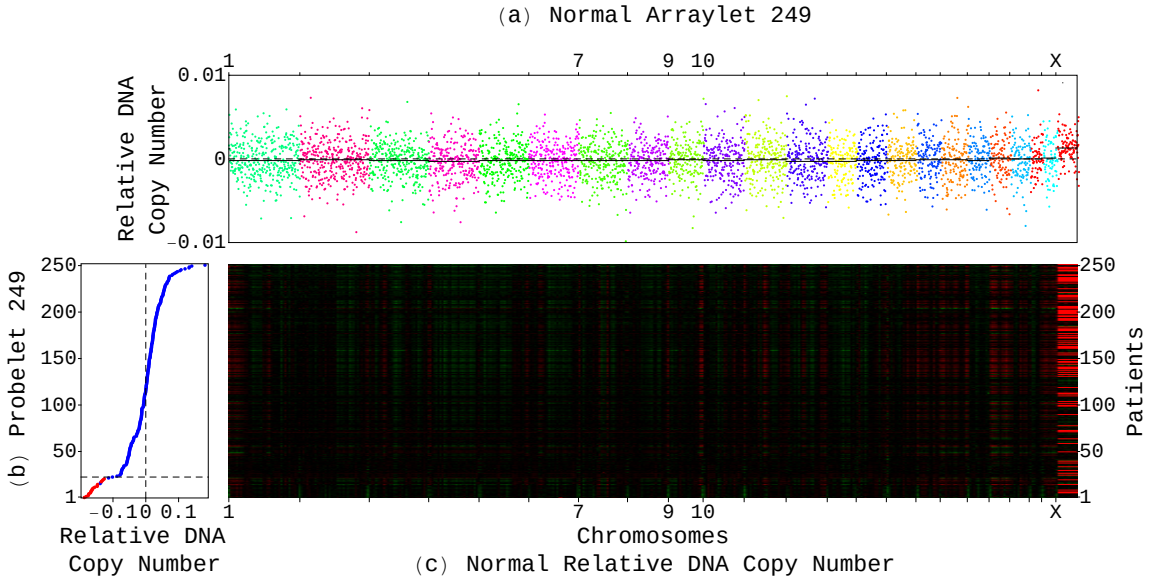


Figure 3.4. The 249th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. (a) Plot of the 249th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the 249th probelet describes the corresponding variation across the patients. Copy numbers in this probelet significantly correlate with the tissue batch/hybridization scanner of the normal samples, HMS 8/2331 (red) and other (gray), with the P-values $<10^{-12}$ (Table 2.1 and Figure 2.4d). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 249th probelet and normal arraylet.

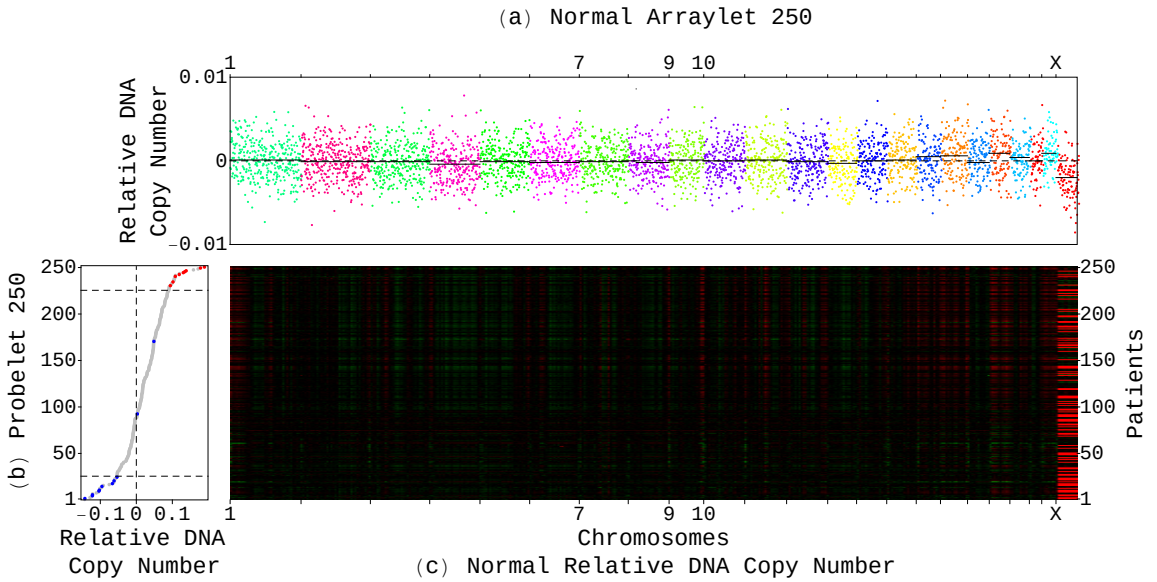


Figure 3.5. The 250th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. (a) Plot of the 250th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the 250th probelet describes the corresponding variation across the patients. Copy numbers in this probelet correlate with the date of hybridization of the normal samples, 4.18.2007 (red), 7.22.2009 (blue) or other (gray), with the P-values $<10^{-3}$ (Table 2.1 and Figure 2.4e). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 250th probelet and normal arraylet.

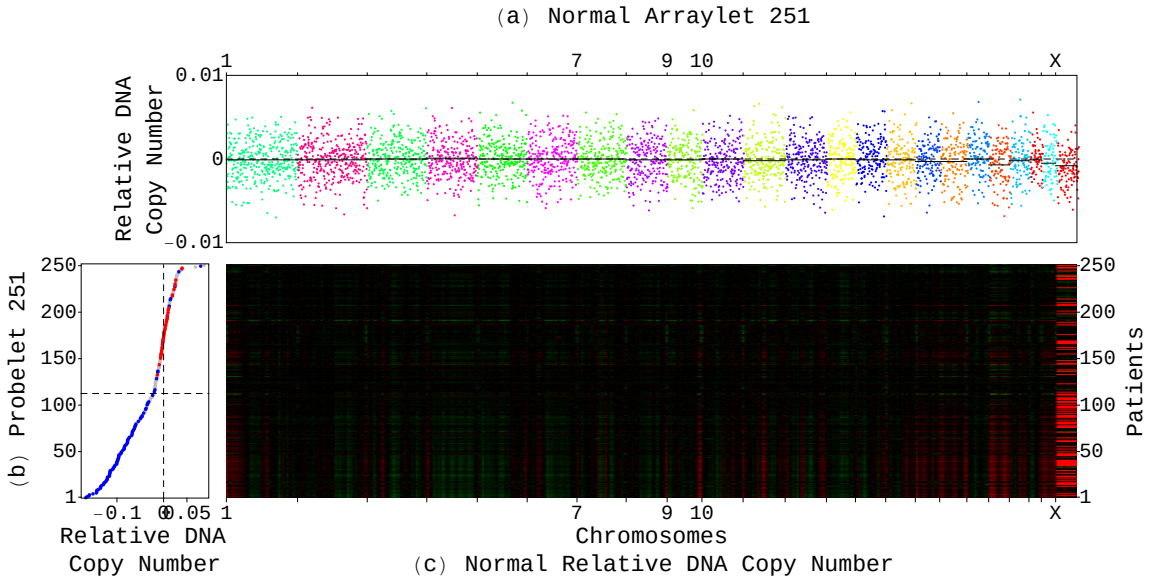


Figure 3.6. The first most normal-exclusive, i.e., 251st probelet and corresponding normal arraylet uncovered by GSVD. (a) Plot of the 251st normal arraylet describes unsegmented chromosomes (black lines), each with copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the first most normal-exclusive probelet, which is also the most significant probelet in the normal dataset (Figure 2.3b), describes the corresponding variation across the patients. Copy numbers in this probelet significantly correlate with the genomic center where the normal samples were hybridized at, HMS (red), MSKCC (blue) or multiple locations (gray), with the P-values $<10^{-13}$ (Table 2.1 and Figure 2.4f). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 251st probelet and normal arraylet.

Figure 3.7. Significant probelets and corresponding tumor and normal arraylets uncovered by GSVD of the patient-matched GBM and normal aCGH profiles. (a) Plot of the second tumor arraylet describes a global pattern of tumor-exclusive co-occurring CNAs across the tumor probes. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. Segments (black lines) identified by circular binary segmentation (CBS) [28, 29] include most known GBM-associated focal CNAs (black), e.g., *EGFR* amplification. CNAs previously unrecognized in GBM (red) include an amplification of a segment containing the biochemically putative drug target-encoding *TLK2*. (b) Plot of the second most tumor-exclusive probelet, which is also the most significant probelet in the tumor dataset (Figure 2.3a), describes the corresponding variation across the patients. The patients are ordered and classified according to each patient's relative copy number in this probelet. There are 227 patients (blue) with high (>0.02) and 23 patients (red) with low, approximately zero, numbers in the second probelet. One patient (gray) remains unclassified with a large negative (<-0.02) number. This classification significantly correlates with GBM survival times (Figure 3.8a and Table 3.1). (c) Raster display of the tumor dataset, with relative gain (red), no change (black) and loss (green) of DNA copy numbers, shows the correspondence between the GBM profiles and the second probelet and tumor arraylet. Chromosome 7 gain and losses of chromosomes 9p and 10, which are dominant in the second tumor arraylet (Figure 3.7a), are negligible in the patients with low copy numbers in the second probelet, but distinct in the remaining patients (Figure 3.7b). This illustrates that the copy numbers listed in the second probelet correspond to the weights of the second tumor arraylet in the GBM profiles of the patients. (d) Plot of the 246th normal arraylet describes an X chromosome-exclusive amplification across the normal probes. (e) Plot of the 246th probelet, which is approximately common to both the normal and tumor datasets, and is the second most significant in the normal dataset (Figure 2.3b), describes the corresponding copy-number amplification in the female (red) relative to the male (blue) patients. Classification of the patients by the 246th probelet agrees with the copy-number gender assignments (Table 3.1 and Figure 2.5), also for three patients with missing TCGA gender annotations and three additional patients with conflicting TCGA annotations and copy-number gender assignments. (f) Raster display of the normal dataset shows the correspondence between the normal profiles and the 246th probelet and normal arraylet. X chromosome amplification, which is dominant in the 246th normal arraylet (Figure 3.7d), is distinct in the female but nonexisting in the male patients (Figure 3.7e). Note also that although the tumor samples exhibit female-specific X chromosome amplification (Figure 3.7c), the second tumor arraylet (Figure 3.7a) exhibits an unsegmented X chromosome copy-number distribution, that is approximately centered at zero with a relatively small width.

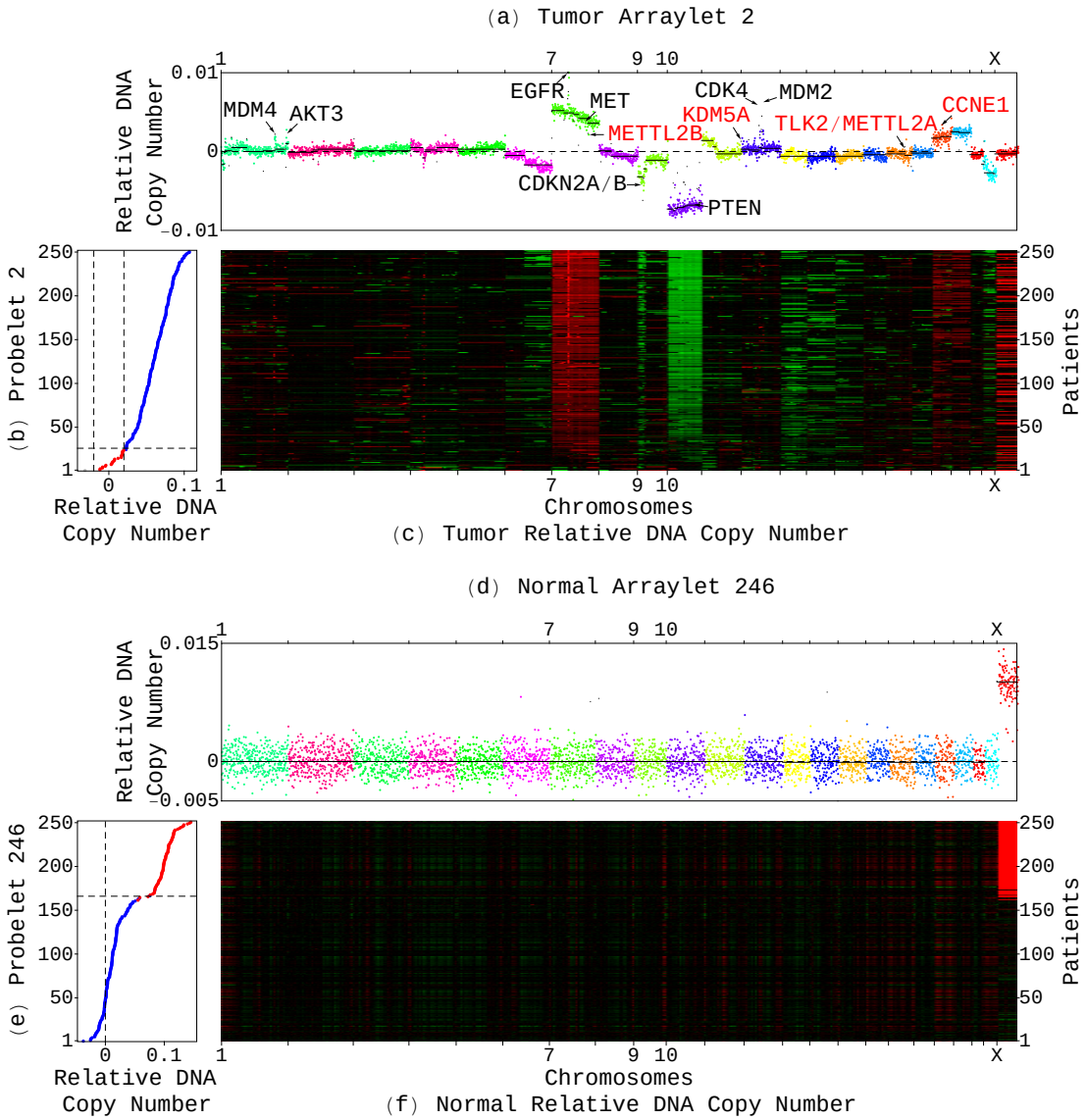
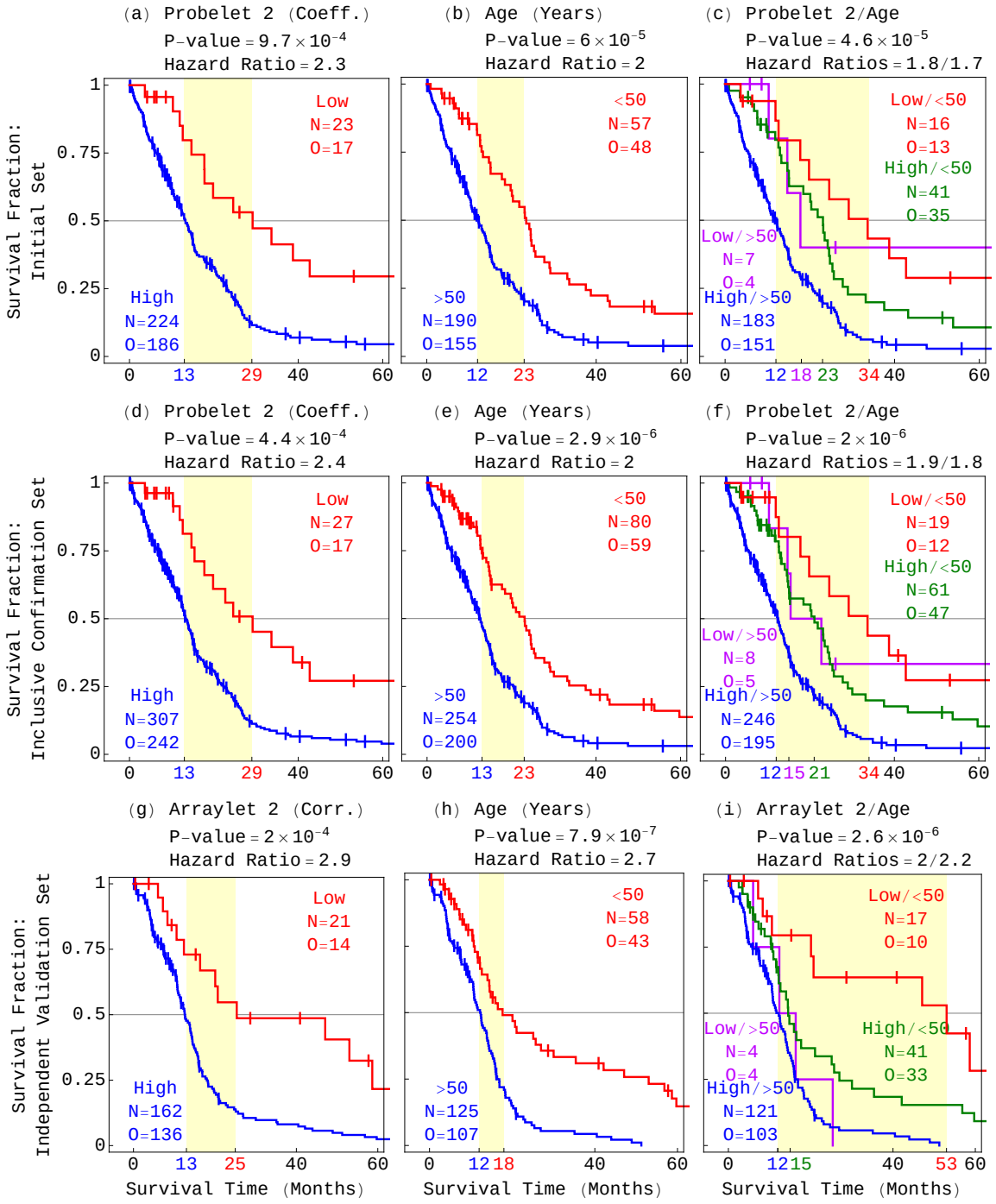

(a) Tumor Arraylet 2
Relative DNA Copy Number
0.01
0
−0.01
1
7
9 10
X
MDM4
AKT3
EGFR
MET
METTL2B
CDK4
MDM2
KDM5A
TLK2/METTL2A
CCNE1
CDKN2A/B
PTEN
(b) Probelet 2
250
200
150
100
50
1
0
0.1
Relative DNA Copy Number
Patients
Chromosomes
(c) Tumor Relative DNA Copy Number
(d) Normal Arraylet 246
0.015
0
−0.005
(e) Probelet 246
(f) Normal Relative DNA Copy Number

Figure 3.8. Survival analyses of the three sets of patients classified by GSVD, age at diagnosis or both. (*a*) Kaplan-Meier (KM) [47] curves for the 247 patients with TCGA annotations in the initial set of 251 patients, classified by copy numbers in the second probelet, which is computed by GSVD for the 251 patients, show a median survival time difference of ∼16 months, with the corresponding log-rank test P-value $< 10^{-3}$. The univariate Cox [48] proportional hazard ratio is 2.3, with a P-value $< 10^{-2}$ (Table 3.1), meaning that high relative copy numbers in the second probelet confer more than twice the hazard of low numbers. The P-values are calculated without adjusting for multiple comparisons [49]. (*b*) Survival analyses of the 247 patients classified by age, i.e., >50 or <50 years old at diagnosis, show that the prognostic contribution of age, with a KM median survival time difference of ∼11 months and a univariate Cox hazard ratio of 2, is comparable to that of GSVD. (*c*) Survival analyses of the 247 patients classified by both GSVD and age, show similar multivariate Cox hazard ratios, of 1.8 and 1.7, that do not differ significantly from the corresponding univariate hazard ratios, of 2.3 and 2, respectively. This means that GSVD and age are independent prognostic predictors. With a KM median survival time difference of ∼22 months, GSVD and age combined make a better predictor than age alone. (*d*) Survival analyses of the 334 patients with TCGA annotations and a GSVD classification in the inclusive confirmation set of 344 patients, classified by copy numbers in the second probelet, which is computed by GSVD for the 344 patients, show a KM median survival time difference of ∼16 months and a univariate hazard ratio of 2.4, and confirm the survival analyses of the initial set of 251 patients. (*e*) Survival analyses of the 334 patients classified by age confirm that the prognostic contribution of age, with a KM median survival time difference of ∼10 months and a univariate hazard ratio of 2, is comparable to that of GSVD. (*f*) Survival analyses of the 334 patients classified by both GSVD and age, show similar multivariate Cox hazard ratios, of 1.9 and 1.8, that do not differ significantly from the corresponding univariate hazard ratios, and a KM median survival time difference of ∼22 months, with the corresponding log-rank test P-value $< 10^{-5}$. This confirms that the prognostic contribution of GSVD is independent of age, and that combined with age, GSVD makes a better predictor than age alone. (*g*) Survival analyses of the 183 patients with a GSVD classification in the independent validation set of 184 patients, classified by correlations of each patient's GBM profile with the second tumor arraylet, which is computed by GSVD for the 251 patients, show a KM median survival time difference of ∼12 months and a univariate hazard ratio of 2.9, and validate the survival analyses of the initial set of 251 patients. (*h*) Survival analyses of the 183 patients classified by age validate that the prognostic contribution of age is comparable to that of GSVD. (*i*) Survival analyses of the 183 patients classified by both GSVD and age, show similar multivariate Cox hazard ratios, of 2 and 2.2, and a KM median survival time difference of ∼41 months, with the corresponding log-rank test P-value $< 10^{-5}$. This validates that the prognostic contribution of GSVD is independent of age, and that combined with age, GSVD makes a better predictor than age alone, also for patients with measured GBM aCGH profiles in the absence of matched normal profiles.


(a) Probelet 2 (Coeff.)
P-value = 9.7 × 10^-4
Hazard Ratio = 2.3
Low
N=23
O=17
High
N=224
O=186
(b) Age (Years)
P-value = 6 × 10^-5
Hazard Ratio = 2
<50
N=57
O=48
>50
N=190
O=155
(c) Probelet 2/Age
P-value = 4.6 × 10^-5
Hazard Ratios = 1.8/1.7
Low/<50
N=16
O=13
High/<50
N=41
O=35
Low/>50
N=7
O=4
High/>50
N=183
O=151
Survival Fraction:
Initial Set
(d) Probelet 2 (Coeff.)
P-value = 4.4 × 10^-4
Hazard Ratio = 2.4
Low
N=27
O=17
High
N=307
O=242
(e) Age (Years)
P-value = 2.9 × 10^-6
Hazard Ratio = 2
<50
N=80
O=59
>50
N=254
O=200
(f) Probelet 2/Age
P-value = 2 × 10^-6
Hazard Ratios = 1.9/1.8
Low/<50
N=19
O=12
High/<50
N=61
O=47
Low/>50
N=8
O=5
High/>50
N=246
O=195
Survival Fraction:
Inclusive Confirmation Set
(g) Arraylet 2 (Corr.)
P-value = 2 × 10^-4
Hazard Ratio = 2.9
Low
N=21
O=14
High
N=162
O=136
(h) Age (Years)
P-value = 7.9 × 10^-7
Hazard Ratio = 2.7
<50
N=58
O=43
>50
N=125
O=107
(i) Arraylet 2/Age
P-value = 2.6 × 10^-6
Hazard Ratios = 2/2.2
Low/<50
N=17
O=10
High/<50
N=41
O=33
Low/>50
N=4
O=4
High/>50
N=121
O=103
Survival Fraction:
Independent Validation Set
Survival Time (Months)
Survival Time (Months)
Survival Time (Months)

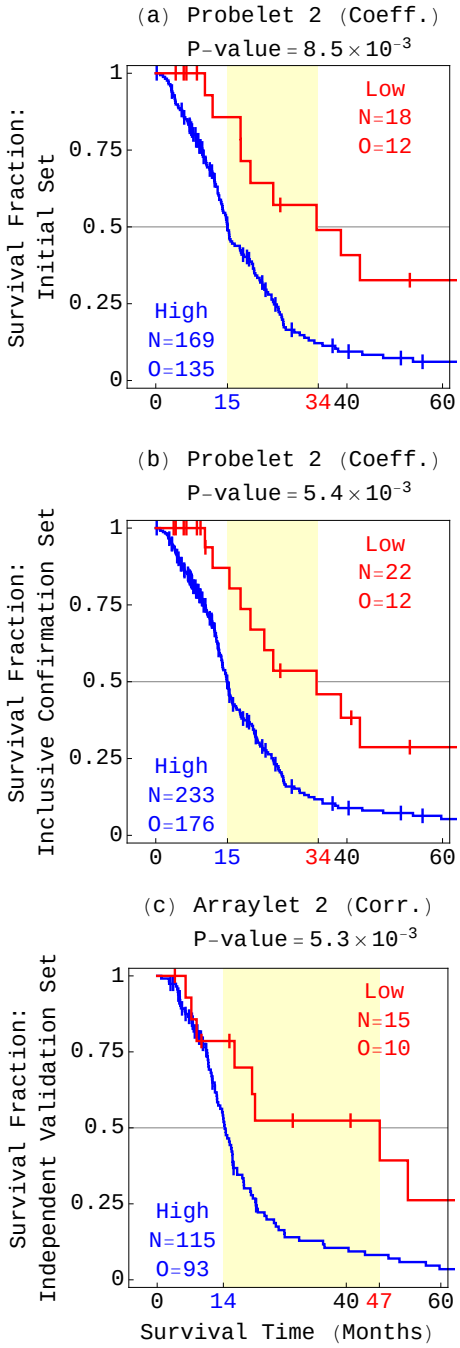


Figure 3.9. Kaplan-Meier (KM) survival analyses of only the chemotherapy patients from the three sets classified by GSVD.

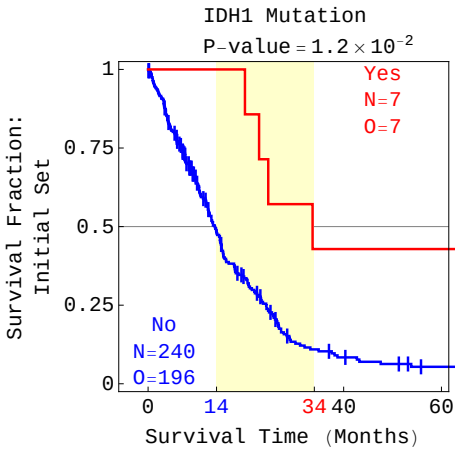


Figure 3.10. KM survival analysis of the initial set of 251 patients classified by a mutation in the gene *IDH1*.

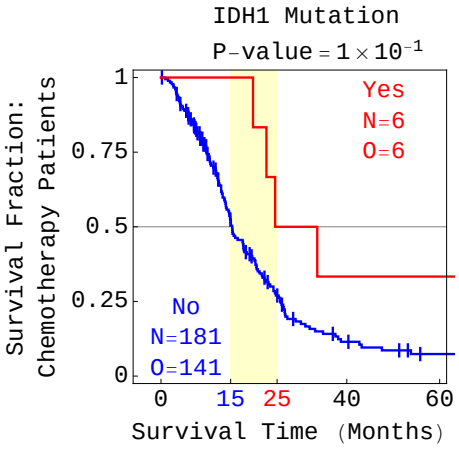


Figure 3.11. KM survival analysis of only the chemotherapy patients in the initial set, classified by a mutation in *IDH1*.

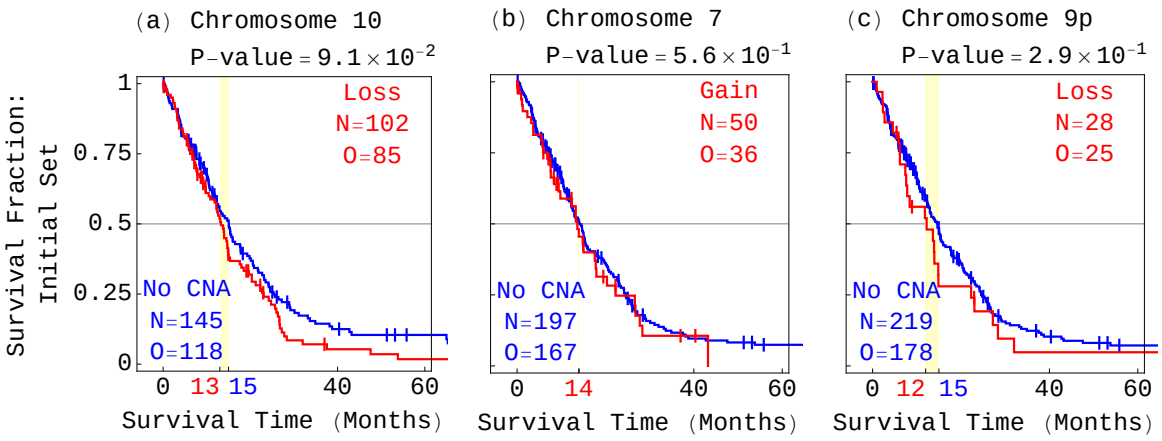
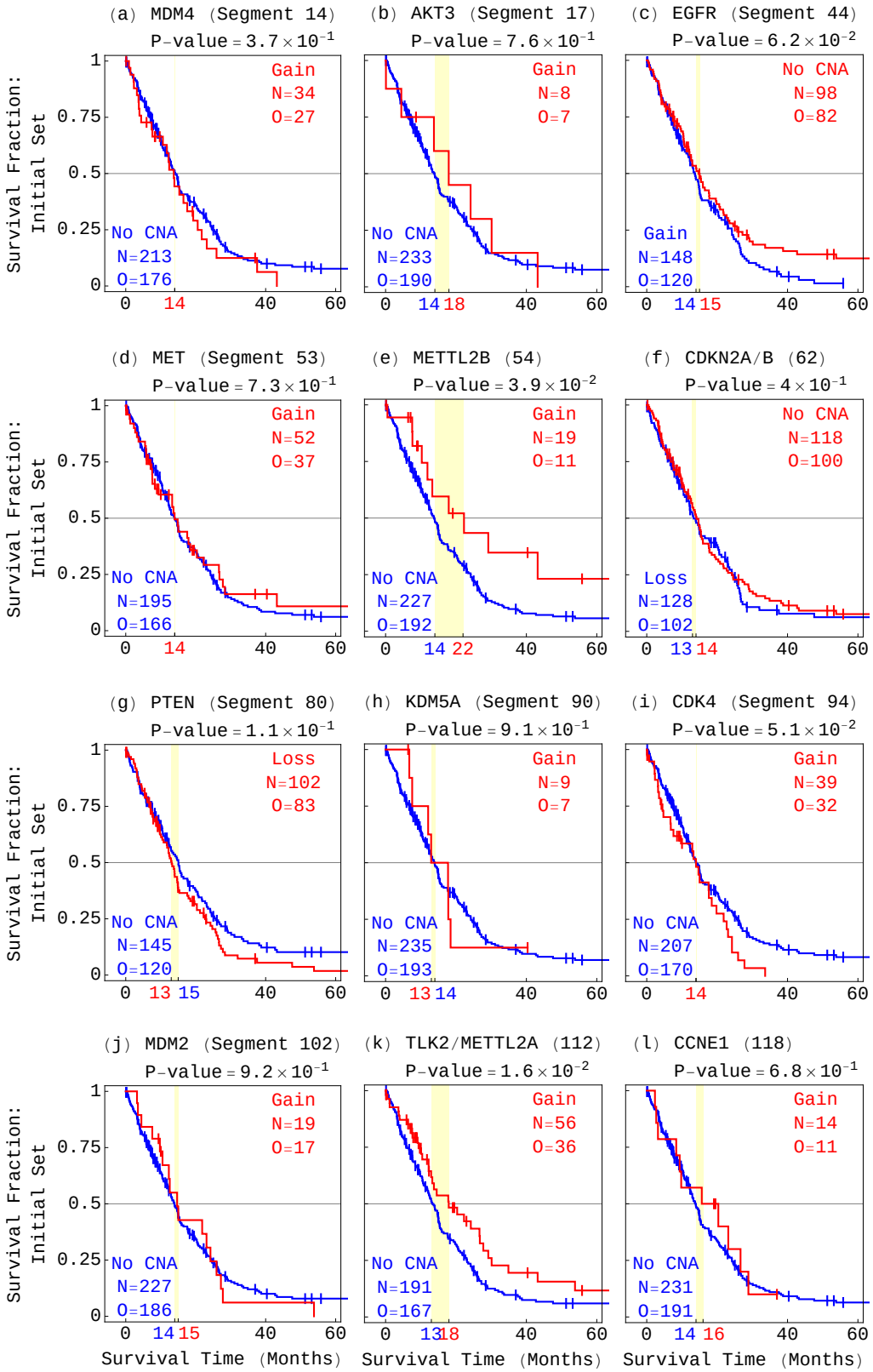


Figure 3.12. KM survival analyses of the initial set of 251 patients classified by GBM-associated chromosome number changes. (a) Analysis of the 247 patients with TCGA annotations in the initial set of 251 patients, classified by number changes in chromosome 10, shows almost overlapping Kaplan-Meier (KM) [36] curves with a KM median survival time difference of $\sim$2 months, and a corresponding log-rank test P-value $\sim 10^{-1}$, meaning that chromosome 10 loss, frequently observed in GBM, is a poor predictor of GBM patients' survival. (b) KM survival analysis of the 247 patients classified by number changes in chromosome 7 shows almost overlapping KM curves with a KM median survival time difference of $<$one month, and a corresponding log-rank test P-value $>5\times 10^{-1}$, meaning that chromosome 7 gain is a poor predictor of GBM survival. (c) KM survival analysis of the 247 patients classified by number changes in chromosome 9p shows a KM median survival time difference of $\sim$3 months, and a log-rank test P-value $>10^{-1}$, meaning that chromosome 9p loss is a poor predictor of GBM survival.

Figure 3.13. KM survival analyses of the initial set of 251 patients classified by copy number changes in selected segments containing GBM-associated genes or genes previously unrecognized in GBM. In the KM survival analyses of the groups of patients with either a CNA or no CNA in either one of the 130 segments identified by the global pattern, i.e., the second tumor-exclusive arraylet (Dataset S3), log-rank test P-values $<5\times10^{-2}$ are calculated for only 12 of the classifications. Of these, only six correspond to a KM median survival time difference that is $\gtrsim$5 months, approximately a third of the $\sim$16 months difference observed for the GSVD classification. One of these segments contains the genes *TLK2* and *METTL2A*, previously unrecognized in GBM. The KM median survival time we calculate for the 56 patients with *TLK2* amplification is $\sim$5 months longer than that for the remaining patients. This suggests that drug-targeting the kinase and/or the methyltransferase-like protein that *TLK2* and *METTL2A* encode, respectively, may affect not only the pathogenesis but also the prognosis of GBM.


(a) MDM4 (Segment 14)
P-value = 3.7 × 10^-1
Gain N=34 O=27
No CNA N=213 O=176
(b) AKT3 (Segment 17)
P-value = 7.6 × 10^-1
Gain N=8 O=7
No CNA N=233 O=190
(c) EGFR (Segment 44)
P-value = 6.2 × 10^-2
No CNA N=98 O=82
Gain N=148 O=120
(d) MET (Segment 53)
P-value = 7.3 × 10^-1
Gain N=52 O=37
No CNA N=195 O=166
(e) METTL2B (54)
P-value = 3.9 × 10^-2
Gain N=19 O=11
No CNA N=227 O=192
(f) CDKN2A/B (62)
P-value = 4 × 10^-1
No CNA N=118 O=100
Loss N=128 O=102
(g) PTEN (Segment 80)
P-value = 1.1 × 10^-1
Loss N=102 O=83
No CNA N=145 O=120
(h) KDM5A (Segment 90)
P-value = 9.1 × 10^-1
Gain N=9 O=7
No CNA N=235 O=193
(i) CDK4 (Segment 94)
P-value = 5.1 × 10^-2
Gain N=39 O=32
No CNA N=207 O=170
(j) MDM2 (Segment 102)
P-value = 9.2 × 10^-1
Gain N=19 O=17
No CNA N=227 O=186
(k) TLK2/METTL2A (112)
P-value = 1.6 × 10^-2
Gain N=56 O=36
No CNA N=191 O=167
(l) CCNE1 (118)
P-value = 6.8 × 10^-1
Gain N=14 O=11
No CNA N=231 O=191
Survival Fraction: Initial Set
Survival Time (Months)

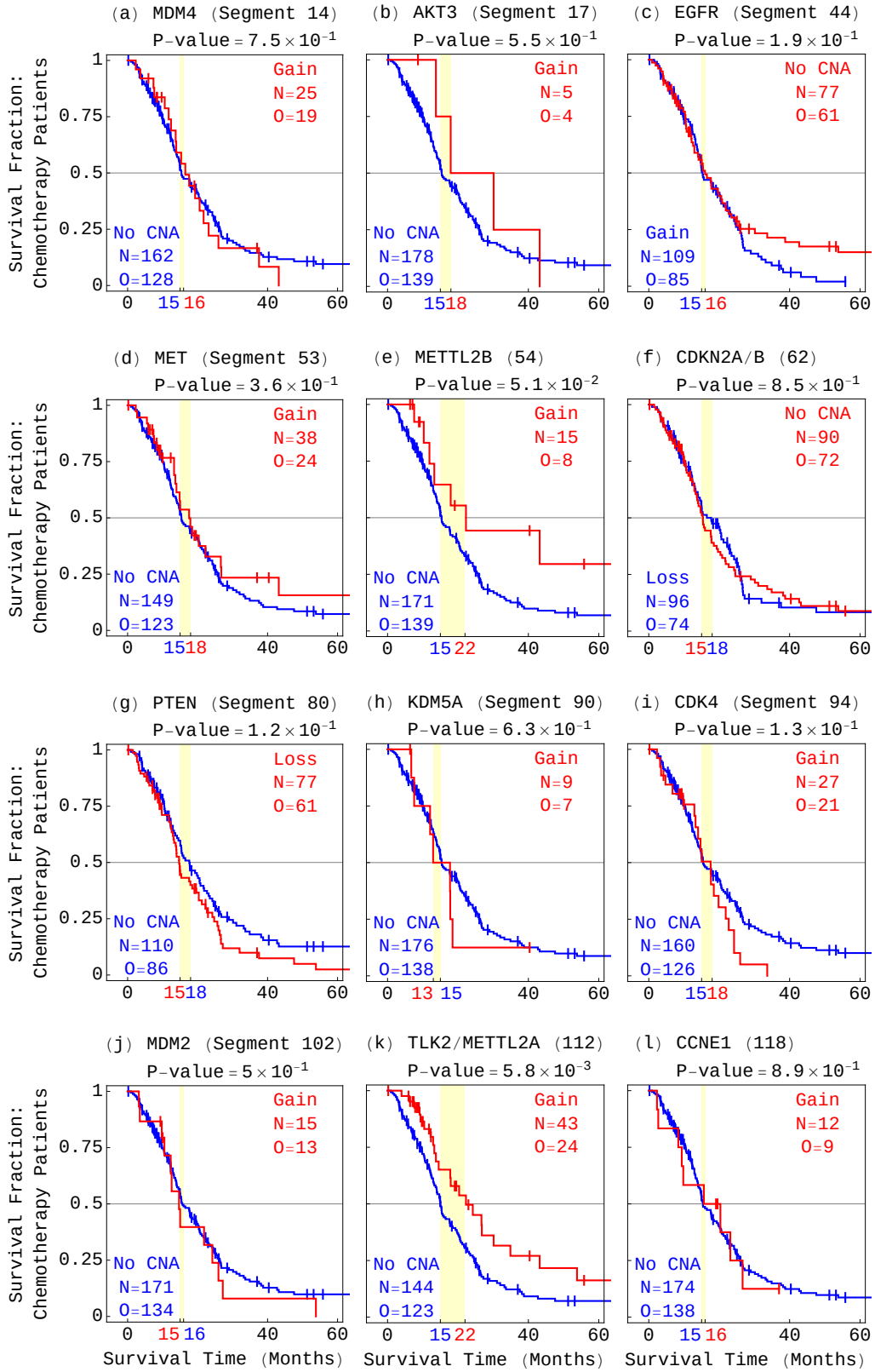


Figure 3.14. KM survival analyses of only the chemotherapy patients in the initial set classified by copy number changes in selected segments.

Figure 3.15. Survival analyses of the patients from the three sets classified by chemotherapy alone or GSVD and chemotherapy both. (a) KM and Cox survival analyses of the 236 patients with TCGA chemotherapy annotations in the initial set of 251 patients, classified by chemotherapy, show that lack of chemotherapy, with a KM median survival time difference of ∼10 months and a univariate hazard ratio of 2.6 (Table 3.2), confers more than twice the hazard of chemotherapy. (b) Survival analyses of the 236 patients classified by both GSVD and chemotherapy, show similar multivariate Cox hazard ratios, of 3 and 3.1, respectively. This means that GSVD and chemotherapy are independent prognostic predictors. With a KM median survival time difference of ∼30 months, GSVD and chemotherapy combined make a better predictor than chemotherapy alone. (c) Survival analyses of the 317 patients with TCGA chemotherapy annotations in the inclusive confirmation set of 344 patients, classified by chemotherapy, show a KM median survival time difference of ∼11 months and a univariate hazard ratio of 2.7, and confirm the survival analyses of the initial set of 251 patients. (d) Survival analyses of the 317 patients classified by both GSVD and chemotherapy show similar multivariate Cox hazard ratios, of 3.1 and 3.2, and a KM median survival time difference of ∼30 months, with the corresponding log-rank test P-value $<10^{-17}$. This confirms that the prognostic contribution of GSVD is independent of chemotherapy, and that combined with chemotherapy, GSVD makes a better predictor than chemotherapy alone. (e) Survival analyses of the 154 patients with TCGA chemotherapy annotations in the independent validation set of 184 patients, classified by chemotherapy, show a KM median survival time difference of ∼11 months and a univariate hazard ratio of 2.2, and validate the survival analyses of the initial set of 251 patients. (f) Survival analyses of the 154 patients classified by both GSVD and chemotherapy, show similar multivariate Cox hazard ratios, of 3.3 and 2.7, and a KM median survival time difference of ∼43 months. This validates that the prognostic contribution of GSVD is independent of chemotherapy, and that combined with chemotherapy, GSVD makes a better predictor than chemotherapy alone, also for patients with measured GBM aCGH profiles in the absence of matched normal profiles.

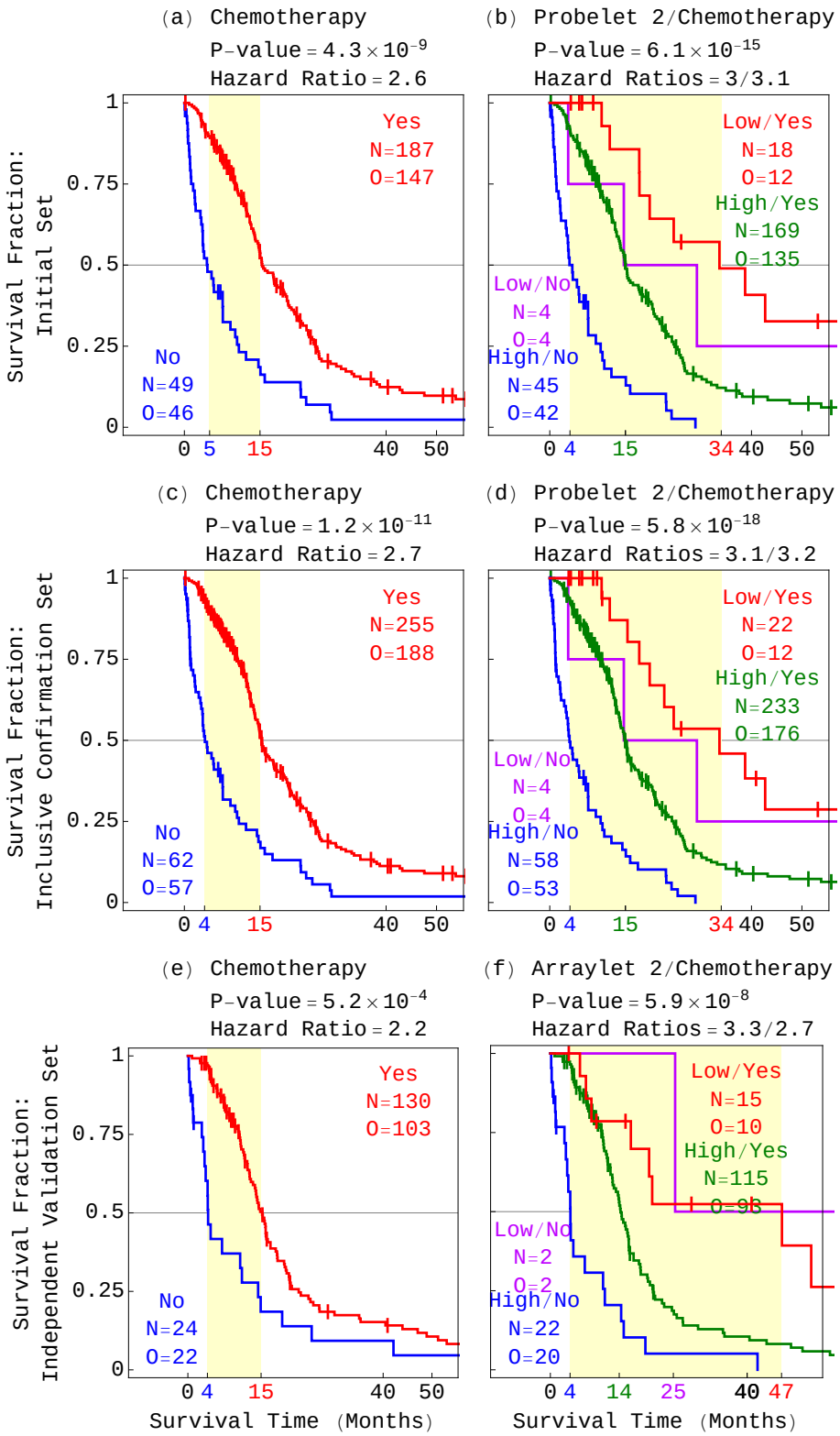

(a) Chemotherapy
P-value = 4.3 × 10^-9
Hazard Ratio = 2.6
Yes
N=187
O=147
No
N=49
O=46
Survival Fraction:
Initial Set
(b) Probelet 2/Chemotherapy
P-value = 6.1 × 10^-15
Hazard Ratios = 3/3.1
Low/Yes
N=18
O=12
High/Yes
N=169
O=135
Low/No
N=4
O=4
High/No
N=45
O=42
(c) Chemotherapy
P-value = 1.2 × 10^-11
Hazard Ratio = 2.7
Yes
N=255
O=188
No
N=62
O=57
Survival Fraction:
Inclusive Confirmation Set
(d) Probelet 2/Chemotherapy
P-value = 5.8 × 10^-18
Hazard Ratios = 3.1/3.2
Low/Yes
N=22
O=12
High/Yes
N=233
O=176
Low/No
N=4
O=4
High/No
N=58
O=53
(e) Chemotherapy
P-value = 5.2 × 10^-4
Hazard Ratio = 2.2
Yes
N=130
O=103
No
N=24
O=22
Survival Fraction:
Independent Validation Set
Survival Time (Months)
(f) Arraylet 2/Chemotherapy
P-value = 5.9 × 10^-8
Hazard Ratios = 3.3/2.7
Low/Yes
N=15
O=10
High/Yes
N=115
O=98
Low/No
N=2
O=2
High/No
N=22
O=20
Survival Time (Months)

Table 3.1. Cox proportional hazard models of the three sets of patients classified by GSVD, age at diagnosis or both. In each set of patients, the multivariate Cox proportional hazard ratios [37] for GSVD and age are similar and do not differ significantly from the corresponding univariate hazard ratios. This means that GSVD and age are independent prognostic predictors.

Cox Proportional Hazard Model	Predictor	Initial Set		Inclusive Confirmation Set		Independent Validation Set	
		Hazard Ratio	P-value	Hazard Ratio	P-value	Hazard Ratio	P-value
Univariate	GSVD	2.3	1.3×10^{-3}	2.4	6.5×10^{-4}	2.9	3.6×10^{-4}
	Age	2.0	7.9×10^{-5}	2.0	4.3×10^{-6}	2.7	1.7×10^{-6}
Multivariate	GSVD	1.8	2.2×10^{-2}	1.9	1.2×10^{-2}	2.0	2.2×10^{-2}
	Age	1.7	2.0×10^{-3}	1.8	1.0×10^{-4}	2.2	2.0×10^{-4}

Table 3.2. Cox proportional hazard models of the three sets of patients classified by GSVD, chemotherapy or both. In each set of patients, the multivariate Cox proportional hazard ratios for GSVD and chemotherapy are similar and do not differ significantly from the corresponding univariate hazard ratios. This means that GSVD and chemotherapy are independent prognostic predictors. The P-values are calculated without adjusting for multiple comparisons [38].

Cox Proportional Hazard Model	Predictor	Initial Set		Inclusive Confirmation Set		Independent Validation Set	
		Hazard Ratio	P-value	Hazard Ratio	P-value	Hazard Ratio	P-value
Univariate	GSVD	2.4	1.2×10^{-3}	2.4	6.4×10^{-4}	2.8	1.3×10^{-3}
	Chemo	2.6	1.5×10^{-8}	2.7	6.3×10^{-11}	2.2	7.3×10^{-4}
Multivariate	GSVD	3.0	5.2×10^{-5}	3.1	2.5×10^{-5}	3.3	2.3×10^{-4}
	Chemo	3.1	7.9×10^{-11}	3.2	1.9×10^{-13}	2.7	3.0×10^{-5}

CHAPTER 4

DISCUSSION

Previously, we showed that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent experimental or biological reality [16, 22]. The variables, subspaces of significant patterns that are common to both or exclusive to either one of the datasets, correlate with cellular programs that are conserved in both or unique to either one of the organisms, respectively. The operation of reconstruction in the subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [14] verify a computationally predicted genome-wide mode of regulation [34, 35, 50], and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

Recently, we mathematically defined a higher-order GSVD (HO GSVD) for more than two large-scale matrices with different row dimensions and the same column dimension [17]. We proved that this novel HO GSVD extends to higher orders almost all of the mathematical properties of the GSVD. We showed, comparing global mRNA expression from the three disparate organisms *S. pombe*, *S. cerevisiae* and human, that the HO GSVD provides a sequence-independent comparative framework for more than two genomic datasets, where the variables and operations represent experimental or biological reality. The approximately common HO GSVD subspace represents biological similarity among the organisms. Simultaneous reconstruction in the common subspace removes the experimental artifacts, which are dissimilar, from the datasets.

We now show that also in probe-independent comparison of aCGH data from patient-matched tumor and normal samples, the mathematical variables of the GSVD, i.e., shared probelets and the corresponding tumor- and normal-specific arraylets, represent experi-

mental or biological reality. Probelets that are mathematically significant in both datasets, correspond to normal arraylets representing copy-number variations (CNVs) in the normal human genome that are conserved in the tumor genome (e.g., female-specific X chromosome amplification) and are represented by the corresponding tumor arraylets. Probelets that are mathematically significant in the normal but not the tumor dataset represent experimental variations that exclusively affect the normal dataset. Similarly, some probelets that are mathematically significant in the tumor but not the normal dataset represent experimental variations that exclusively affect the tumor dataset.

We find that the mathematically second most tumor-exclusive probelet, which is also the mathematically most significant probelet in the tumor dataset, is statistically correlated, possibly biologically coordinated with GBM patients' survival and response to chemotherapy. The corresponding tumor arraylet describes a global pattern of tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs, including the biochemically putative drug target-encoding *TLK2* [18]. We find that a negligible weight of the second tumor arraylet in a patient's GBM aCGH profile, mathematically defined by either the corresponding copy number in the second probelet, or by the correlation of the GBM profile with the second arraylet, is indicative of a significantly longer GBM survival time. This GSVD comparative modeling of aCGH data from patient-matched tumor and normal samples, therefore, draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.

We confirm our results with GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set. We validate the prognostic contribution of the pattern with GSVD classification of the GBM profiles of a set of patients that is independent of both the initial set and the inclusive confirmation set [19].

Additional possible applications of the GSVD (and also the HO GSVD) in personalized medicine include comparisons of multiple patient-matched datasets, each corresponding to either (*i*) a set of large-scale molecular biological profiles (such as DNA copy numbers) acquired by a high-throughput technology (such as DNA microarrays) from the same tissue type (such as tumor or normal); or (*ii*) a set of biomedical images or signals; or (*iii*) a set of anatomical or clinical pathology test results or phenotypical observations (such as age). GSVD comparisons can uncover the relations and possibly even causal coordinations between these different recorded aspects of the same medical phenomenon.

GSVD comparisons can be used to determine a single patient's medical status in relation to all the other patients in the set, and inform the patient's diagnosis, prognosis and treatment.

APPENDIX

SUPPORTING INFORMATION

Mathematica Notebook S1. Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles. A Mathematica 8.0.1 code file, executable by Mathematica 8.0.1 and readable by Mathematica Player, freely available at
http://www.wolfram.com/products/player/.
(NB)

Mathematica Notebook S2. Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles. A PDF format file, readable by Adobe Acrobat Reader.
(PDF)

Dataset S1. Initial set of 251 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing The Cancer Genome Atlas (TCGA) [5] annotations of the initial set of 251 patients and the corresponding normal and tumor samples. The tumor and normal profiles of the initial set of 251 patients, in tab-delimited text format files, tabulating $\log_2$ relative copy number variation across 212,696 and 211,227 tumor and normal probes, respectively, are available at http://www.alterlab.org/GBM_prognosis/.
(TXT)

Dataset S2. Segments of the significant tumor and normal arraylets, computed by GSVD for the initial set of 251 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, tabulating segments identified by circular binary segmentation (CBS) [28, 29].
(TXT)

Dataset S3. Segments of the second tumor arraylet, computed by GSVD for the initial set of 251 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, tabulating, for each of the 130 CBS segments of the second tumor arraylet, the segment's coordinates, the CBS P-value, and the log-rank test P-value corresponding to the Kaplan-Meier (KM) survival analysis of the initial set of 251 patients classified by either a gain or a loss of this segment.
(TXT)

Dataset S4. Inclusive confirmation set of 344 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the TCGA annotations of the inclusive confirmation set of 344 patients. The tumor and normal profiles of the inclusive confirmation set of 344 patients, in tab-delimited text format files, tabulating $\log_2$ relative copy number variation across 200,139 and 198,342 tumor and normal probes, respectively, are available at http://www.alterlab.org/GBM_prognosis/.
(TXT)

Dataset S5. Independent validation set of 184 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the TCGA annotations of the independent validation set of 184 patients. Tumor profiles of the independent validation set of 184 patients, in a tab-delimited text format file, tabulating $\log_2$ relative copy number variation across 212,696 autosomal and X chromosome probes, is available at http://www.alterlab.org/GBM_prognosis/.
(TXT)

REFERENCES

[1] Yang D, Khan S, Sun Y, Hess K, Shmulevich I, Sood AK, Zhang W, "Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer," *JAMA*, vol. 306, pp. 1557–1565, 2011.

[2] Purow B, Schiff D, "Advances in the genetics of glioblastoma: are we reaching critical mass?" *Nat Rev Neurol*, vol. 5, pp. 419–426, 2009.

[3] Wiltshire RN, Rasheed BK, Friedman HS, Friedman AH, Bigner SH, "Comparative genetic patterns of glioblastoma multiforme: potential diagnostic tool for tumor classification," *Neuro Oncol*, vol. 2, pp. 164–173, 2000.

[4] Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, et al., "Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma," *Cancer Res*, vol. 65, pp. 1678–1686, 2005.

[5] Cancer Genome Atlas Research Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[6] Mischel PS, Shai R, Shi T, Horvath S, Lu KV, et al., "Identification of molecular subtypes of glioblastoma by gene expression profiling," *Oncogene*, vol. 22, pp. 2361–2673, 2003.

[7] Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, et al., "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, pp. 98–110, 2010.

[8] Colman H, Zhang L, Sulman EP, McDonald JM, Shooshtari NL, et al., "A multigene predictor of outcome in glioblastoma," *Neuro Oncol*, vol. 12, pp. 49–57, 2010.

[9] Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, et al., "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, vol. 17, pp. 510–522, 2010.

[10] Curran WJ Jr, Scott CB, Horton J, Nelson JS, Weinstein AS, et al., "Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials," *J Natl Cancer Inst*, vol. 85, pp. 704–710, 1993.

[11] Gorlia T, van den Bent MJ, Hegi ME, Mirimanoff RO, Weller M, et al., "Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3," *Lancet Oncol*, vol. 9, pp. 29–38, 2008.

[12] Kahn SD, "On the future of genomic data," *Science*, vol. 331, pp. 728–729, 2011.

[13] Alter O, Brown PO, Botstein D, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci USA*, vol. 97, pp. 10 101–10 106, 2000, http://dx.doi.org/10.1073/pnas.97.18.10101.

[14] Omberg L, Meyerson JR, Kobayashi K, Drury LS, Diffley JF, et al., "Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression," *Mol Syst Biol*, vol. 5, p. 312, 2009, http://dx.doi.org/10.1038/msb.2009.70.

[15] Golub GH, Van Loan CF, *Matrix Computations*. Baltimore: Johns Hopkins University Press, third edition, 694 p., 1996.

[16] Alter O, Brown PO, Botstein D, "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms," *Proc Natl Acad Sci USA*, vol. 100, pp. 3351–3356, 2003, http://dx.doi.org/10.1073/pnas.0530258100.

[17] Ponnapalli SP, Saunders MA, Van Loan CF, Alter O, "A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms," *PLoS ONE*, vol. 6, p. e28072, 2011, http://dx.doi.org/10.1371/journal.pone.0028072.

[18] Lee CH, Alter O, "Known and novel copy number alterations in GBM and their patterns of co-occurrence are revealed by GSVD comparison of array CGH data from patient-matched normal and tumor TCGA samples," in *60th Annual American Society of Human Genetics (ASHG) Meeting (November 2–6, 2010, Washington, DC)*, 2010.

[19] Alpert BO, Sankaranarayanan P, Lee CH, Alter O, "Glioblastoma multiforme prognosis by using a patient's array CGH tumor profile and a generalized SVD-computed global pattern of copy-number alterations," in *2nd DNA and Genome World Day (April 25–29, 2011, Dalian, China)*, 2011.

[20] Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, et al., "Molecular characterisation of soft tissue tumours: a gene expression study," *Lancet*, vol. 359, pp. 1301–1307, 2002.

[21] Alter O, Brown PO, Botstein D, "Processing and modeling genome-wide expression data using singular value decomposition," *SPIE*, vol. 4266, pp. 171–186, 2001, http://dx.doi.org/10.1117/12.427986.

[22] Alter O, "Discovery of principles of nature from mathematical modeling of DNA microarray data," *Proc Natl Acad Sci USA*, vol. 103, pp. 16 063–16 064, 2006, http://dx.doi.org/10.1073/pnas.0607650103.

[23] Muralidhara C, Gross AM, Gutell RR, Alter O, "Tensor decomposition reveals concurrent evolutionary convergences and divergences and correlations with structural motifs in ribosomal RNA," *PLoS One*, vol. 6, p. e18768, 2011, http://dx.doi.org/10.1371/journal.pone.0018768.

[24] Alter O, Golub GH, "Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations," *Proc Natl Acad Sci USA*, vol. 102, pp. 17 559–17 564, 2005, http://dx.doi.org/10.1073/pnas.0509033102.

[25] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM, "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281–285, 1999.

[26] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al., "The human genome browser at UCSC," *Genome Res*, vol. 12, pp. 996–1006, 2002.

[27] Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al., "The UCSC Genome Browser database: update 2011," *Nucleic Acids Res*, vol. 39, pp. D876–D882, 2011.

[28] Olshen AB, Venkatraman ES, Lucito R, Wigler M, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, pp. 557–572, 2004.

[29] Venkatraman ES, Olshen AB, "A faster circular binary segmentation algorithm for the analysis of array CGH data," *Bioinformatics*, vol. 23, pp. 657–663, 2007.

[30] Heidenblad M, Lindgren D, Veltman JA, Jonson T, Mahlamäki EH, et al., "Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications," *Oncogene*, vol. 24, pp. 1794–1801, 2005.

[31] Wang Q, Diskin S, Rappaport E, Attiyeh E, Mosse Y, et al., "Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number," *Cancer Res*, vol. 66, pp. 6050–6062, 2006.

[32] Silljé HH, Takahashi K, Tanaka K, Van Houwe G, Nigg EA, "Mammalian homologues of the plant Tousled gene code for cell-cycle-regulated kinases with maximal activities linked to ongoing DNA replication," *EMBO J*, vol. 18, pp. 5691–5702, 1999.

[33] Hopkins AL, Groom CR, "The druggable genome," *Nat Rev Drug Discov*, vol. 1, pp. 727–730, 2002.

[34] Alter O, Golub GH, "Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription," *Proc Natl Acad Sci USA*, vol. 101, pp. 16 577–16 582, 2004, http://dx.doi.org/10.1073/pnas.0406767101.

[35] Omberg L, Golub GH, Alter O, "A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies," *Proc Natl Acad Sci USA*, vol. 104, pp. 18 371–18 376, 2007, http://dx.doi.org/10.1073/pnas.0709146104.

[36] Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, et al., "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process," *BMC Cancer*, vol. 7, p. 64, 2007.

[37] Pellegrini M, Cheng JC, Voutila J, Judelson D, Taylor J, et al., "Expression profile of CREB knockdown in myeloid leukemia cells," *BMC Cancer*, vol. 8, p. 264, 2008.

[38] Millour M, Charbonnel C, Magrangeas F, Minvielle S, Campion L, et al., "Gene expression profiles discriminate between pathological complete response and resistance to neoadjuvant FEC100 in breast cancer," *Cancer Genomics Proteomics*, vol. 3, pp. 89–95, 2006.

[39] Snijders AM, Nowee ME, Fridlyand J, Piek JM, Dorsman JC, et al., "Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in fallopian tube carcinoma," *Oncogene*, vol. 22, pp. 4281–4286, 2003.

[40] Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, et al., "The patterns and dynamics of genomic instability in metastatic pancreatic cancer," *Nature*, vol. 467, pp. 1109–1113, 2010.

[41] Alter O, "Genomic signal processing: From matrix algebra to genetic networks," *Methods Mol Biol*, vol. 377, pp. 17–60, 2007, http://dx.doi.org/10.1007/978-1-59745-390-5_2.

[42] Etemadmoghadam D, George J, Cowin PA, Cullinane C, Kansara M, et al., "Amplicon-dependent CCNE1 expression is critical for clonogenic survival after cisplatin treatment and is correlated with 20q11 gain in ovarian cancer," *PLoS ONE*, vol. 5, p. e15498, 2010.

[43] Defeo-Jones D, Huang PS, Jones RE, Haskell KM, Vuocolo GA, et al., "Cloning of cDNAs for cellular proteins that bind to the retinoblastoma gene product," *Nature*, vol. 352, pp. 251–254, 1991.

[44] Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, et al., "A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations," *Cell*, vol. 141, pp. 69–80, 2010.

[45] Pardridge WM, "The blood-brain barrier: bottleneck in brain drug development," *NeuroRx*, vol. 2, pp. 3–14, 2005.

[46] Hattori Y, Ohta S, Hamada K, Yamada-Okabe H, Kanemura Y, et al., "Identification of a neuron-specific human gene, KIAA1110, that is a guanine nucleotide exchange factor for ARF1," *Biochem Biophys Res Commun*, vol. 364, pp. 737–742, 2007.

[47] Kaplan EL, Meier P, "Nonparametric estimation from incomplete observations," *J Amer Statist Assn*, vol. 53, pp. 457–481, 1958.

[48] Cox DR, "Regression models and life-tables," *J Roy Statist Soc B*, vol. 34, pp. 187–220, 1972.

[49] Rothman KJ, "No adjustments are needed for multiple comparisons," *Epidemiology*, vol. 1, pp. 43–46, 1990.

[50] Alter O, Golub GH, Brown PO, Botstein D, "Novel genome-scale correlation between DNA replication and RNA transcription during the cell cycle in yeast is predicted by data-driven models," *MNBWS*, vol. 15, 2004, http://www.med.miami.edu/mnbws/documents/Alter-.pdf.